Probing the phonetic and phonological knowledge of tones in Mandarin TTS models

Jian Zhu

University of Michigan, Ann Arbor, USA
lingjzhu@umich.edu

Abstract

This study probes the phonetic and phonological knowledge of lexical tones in TTS models through two experiments. Controlled stimuli for testing tonal coarticulation and tone sandhi in Mandarin were fed into Tacotron 2 and WaveGlow to generate speech samples, which were subject to acoustic analysis and human evaluation. Results show that both baseline Tacotron 2 and Tacotron 2 with BERT embeddings capture the surface tonal coarticulation patterns well but fail to consistently apply the Tone-3 sandhi rule to novel sentences. Incorporating pretrained BERT embeddings into Tacotron 2 improves the naturalness and prosody performance, and yields better generalization of Tone-3 sandhi rules to novel complex sentences, although the overall accuracy for Tone-3 sandhi was still low. Given that TTS models do capture some linguistic phenomena, it is argued that they can be used to generate and validate certain linguistic hypotheses. On the other hand, it is also suggested that linguistically informed stimuli should be included in the training and the evaluation of TTS models.

Index Terms: speech synthesis, lexical tones, neural networks

1. Introduction

Recent advancements in end-to-end text-to-speech (TTS) synthesis (i.e., Tacotron 2 [1], TransformerTTS [2], FastSpeech [3]) and waveform generation (e.g., WaveNet [4], SampleRNN [5], WaveGlow [6]) have made it possible to generate extremely natural speech that can sometimes even fool human evaluators. Given such impressive performance, these models have necessarily captured at least some, if not all, statistical regularities in the highly varying speech signals. One is therefore naturally inclined to ask: to what extent are these deep neural networks capturing the statistical regularities and/or irregularities of speech?

While such works are lacking in speech science, researchers in natural language processing (NLP) have investigated the extent to which the behaviors of neural language models are comparable to human syntactic abilities. They evaluate deep language models with controlled psycholinguistic stimuli, showing that at least some aspects of neural networks are similar to human language processing, despite key differences [7, 8, 9]. While these studies focus on whether neural networks learn human-like linguistic abilities, it has been suggested that neural networks can also benefit language research by providing computational models for generating or testing linguistic hypotheses [10]. Recent investigations have applied representation learning to study tones and intonations in several languages [11, 12].

The current work goes one step further to ask: is it possible to exploit these learned parameters to gain a better understanding of speech? TTS models seem to combine the advantages of both experimental and corpus-based approaches. They are trained on many hours of speech and therefore are potentially more generalizable to diverse linguistic patterns. Once a TTS model is trained, it can be used to generate speech samples from texts unseen in the training data. If the quality of generated samples is good enough, TTS models can be employed to generate speech from controlled sentence stimuli, which can be subjected to the same acoustic analysis and statistical testing as in most phonetic experiments. However, before this approach is applied to linguistic research, it is important to understand what phonetic and phonological knowledge these TTS models have learned.

2. The current study

To validate the approach outlined in the introduction, this study assesses TTS models' learned knowledge of tonal coarticulation and Tone-3 sandhi in Mandarin with controlled stimuli. In Mandarin, lexical meanings are distinguished through four contrastive tones (Tone-1: high-level; Tone-2: rising; Tone-3: dipping; Tone-4: falling), though some suggest that a fifth neural tone also exists [13]. Tonal coarticulation refers to the within-category variation of lexical tones induced by neighboring tones, which will not change the tone type. Tone sandhi is the mandatory categorical change of tone types caused by neighboring tones. In Mandarin, the most common example is Tone-3 sandhi, in which Tone-3 changes to Tone-2 if it precedes another Tone-3. This process is automatic and mandatory, although subject to some syntactic constraints.

Tonal coarticulation and tone sandhi represent gradient and categorical tonal variations in Mandarin respectively. Both have been extensively investigated for Mandarin [14, 15, 16, 17], making them suitable probes to investigate the phonetic and phonological knowledge of TTS models and verify the proposed approach.

The current work has the following contributions. This is among the first studies to probe the phonetic and phonological knowledge of TTS models, a step towards better understanding of the inner workings of neural networks from a linguistic perspective. It shows that TTS models can be used in linguistic research to explore certain hypotheses because they capture well the gradient phonetic processes in surface form, though still falling short of learning the underlying categorical phonological rules. Secondly, this work also highlights the importance of designing linguistically informed stimuli to train evaluate TTS models.

3. Methods

In this study, stimuli are fed into TTS models to generate speech samples, which are later analyzed as in linguistic experiments. A speech synthesis model (here, Tacotron 2 [1]) takes textual stimuli as input to predict the corresponding mel-spectrogram, and then the log mel-spectrogram is converted to raw waveform through a vocoder (here, WaveGlow [6]).

3.1. Stimuli

3.1.1. Tonal coarticulation

The goal of this experiment is to examine both anticipatory and carry-over tonal coarticulation in Mandarin. Tones in Mandarin can be characterized by their onset and offset values (Table 1) [14]. Anticipatory coarticulation refers to the influence of the onset value of the upcoming tone on the target tone, whereas carry-over coarticulation refers to the tonal variation due to the offset value of the previous tone. The stimuli were all possible bisyllabic combinations of syllables [ma, mo, mi] with different tones embedded in 6 carrier phrases, resulting in 576 sentences in total. Most of these words were designed to be non-sense to exclude the influences of extraneous factors (e.g., frequency, semantic predictability).

Table 1: *Onset and offset of Mandarin tones.*

		Offset	
		High	Low
Onset	High	Tone 1	Tone 4
	Low	Tone 2	Tone 3

3.1.2. Tone-3 sandhi

This experiment assesses whether TTS models can correctly perform the Tone-3 sandhi rules without explicit supervision. Let σ represent a syllable which can carry different lexical tones ($\sigma_1, \sigma_2, \sigma_3, \sigma_4$) and brackets $[\quad]_w$ represent a word. If two Tone-3 syllables co-occur, the first one will change to Tone-2, that is, $[\sigma_3\sigma_3]_w \rightarrow [\sigma_2\sigma_3]_w$. However, when three Tone-3 syllables co-occur, there are two patterns of tone change depending on the syntactic constituency (left-branching or right-branching).

- $[\sigma_3\sigma_3\sigma_3]_w \rightarrow [\sigma_3[\sigma_2\sigma_3]_w]_w$
  Ex. mi3 lao3 - shu3 'Micky mouse' → mi3 lao2 - shu3
- $[\sigma_3\sigma_3\sigma_3]_w \rightarrow [[\sigma_2\sigma_2]_w\sigma_3]_w$.
  Ex. meng3 - gu3 yu3 'Mongolian' → meng2 - gu2 yu3

In other words, syntactic analysis is required to correctly apply the Tone-3 sandhi rule [17] in phrases. Challenging and uncertain cases arises when Tone-3 sequences occur in more complex syntactic structures. Individual differences also exist, as $[\sigma_3]_w[\sigma_3]_w[\sigma_3]_w$ can be realized as any one of the above forms because of speaker choice.

The stimuli for Tone-3 sandhi were 38 bisyllabic words, 32 trisyllabic and 39 phrases that range from 4 to 19 characters, all of which were in Tone-3. The list was compiled through gathering stimuli from previous studies [16, 17] or online corpora.

3.2. Tacotron 2 with BERT embeddings

The end-to-end speech synthesis model used in this study, Tacotron 2 [1], is a state-of-the-art speech synthesis model. As a sequence-to-sequence recurrent neural network, Tacotron 2 can predict log mel-spectrograms from only raw textual inputs. The alignment between textual inputs and predicted spectrograms are learned through location-sensitive attention mechanism.

The original Tacotron 2 only requires phonemes or characters as input. Recent works [18, 19, 20] have explored using syntactic and semantic representations from Bidirectional Encoder Representations from Transformers (BERT), the state-of-the-art pre-trained language model for textual feature extraction [21], achieving better quality than phoneme/character-only models. In this experiment, two Tacotron 2 models were trained, Tacotron 2 with only Pinyin inputs and Tacotron 2 with Pinyin and BERT embeddings. Given the large number of characters, it is unrealistic to use character-based features in Mandarin TTS. Pinyin (phonemic) features are instead adopted as the input features. The syllable structure of Mandarin is relatively simple, as each Chinese character is represented by either one or two Pinyin symbols, a combination of initials and finals, or onsets and rhymes. No explicit tone sandhi information was supplemented in the grapheme-to-phoneme process, as TTS models were expected to pick up the tone sandhi rules in a purely data-driven approach. The BERT embeddings are extracted for each character with the character-based Chinese BERT [22]. Finally they are upsampled to match the length of the Pinyin symbols of each character. Then the BERT embeddings and phonemic embeddings are concatenated as the input to the subsequent layers (Fig. 1). Except for the inputs and the first convolutional layer that mapped inputs into lower dimensions, both TTS models have the exact same architecture and hyperparameters.

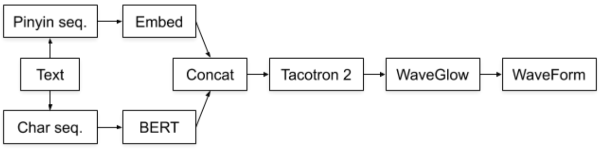

Figure 1: *Workflow of Tacotron 2 + BERT and WaveGlow*

3.3. WaveGlow

WaveGlow is a non-autoregresssive vocoder capable of converting mel-spectrograms to waveforms faster than real time [6]. Previous vocoders such as WaveNet [4] and SampleRNN [5] can generate extremely natural speech waveforms from mel-spectrograms but are painfully slow because of the autoregressive generation process. WaveGlow combines the flow-based generative modeling [23] and WaveNet to achieve non-autoregressive waveform generation, making it possible to massively speed up the generation process while preserving the naturalness of synthesized speech. In this study, a WaveGlow vocoder for Mandarin speech was trained on log mel-spectrograms that match the outputs of Tacotron 2.

4. Experiments

The code, pre-trained models, stimuli and detailed experimental settings are all available online at [https://github.com/lingjzhu/probing-TTS-models].

4.1. Data

The training data for both Tacotron 2 and WaveGlow was the Chinese Standard Mandarin Speech Corpus (CSMSC)[1]. CSMSC has 10,000 recorded sentences read by a female speaker, totaling 12 hours of natural speech with phoneme-level Textgrid annotations and text transcriptions. The corpus was randomly partitioned into non-overlapping training, development and test sets with 9800, 100, 100 sentences respectively.

[1] https://www.data-baker.com/open_source.html

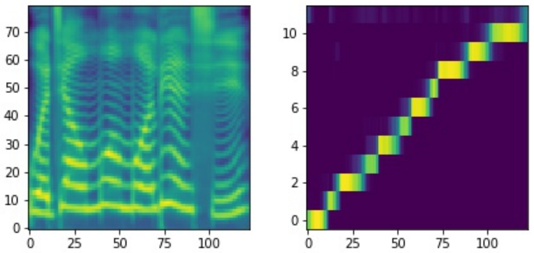

Figure 2: *Predicted log mel-spectrogram (left) and attention alignment plot (right) for* ***wo3 jiao4 mo4 mo4 lian4 xi2*** *(”I asked mo mo to practice”).*

4.2. Model training

4.2.1. Tacotron 2

The Tacotron 2 was based on the NVIDIA's public implementation and the pre-trained model[2]. The model was trained with a batch size of 64 and the Adam optimizer (lr=1e-3). Weights of the Tacotron 2 model were initialized with the pre-trained English model for compatible layers. It is found that weights for English speech synthesis can accelerate the convergence of the Mandarin model. The model began to form diagonal attention alignment in one thousand iterations and produce intelligible speech after only several thousand iterations. The Mandarin Tacotron 2 was trained for 50k iterations on a single RTX 2080Ti. The entire training took less than 2 days to complete.

4.2.2. WaveGlow

All default settings of the original WaveGlow implementation[3] were kept. The batch size was set to 8 and the Adam optimizer with a learning rate of 1e-4 was used. The Mandarin WaveGlow model was also initialized with the weights of the pre-trained English WaveGlow, because transferring weights from a pre-trained model, even in a different language, greatly speeds up the training process. With the warm start, the model produced good-quality speech at only 46k iterations but the final model was trained for 150k iterations. The training process was done on a single RTX 2080Ti GPU, which lasted 3 days.

4.3. Evaluation

Acoustic analysis was carried out to examine the generated speech samples for tonal coarticulation. The time-aligned Textgrid annotations for each stimulus were derived from the attention alignment (Fig. 2) through a simple greedy approach. Attention alignment does not always translate to accurate transcriptions so the generated Textgrids were inspected. Three sentences were excluded due to problematic annotation. The f0 tracking was done using the auto-correlation method through the Parselmouth API of Praat [24].

The rest of the speech samples were assessed by five trained linguists who were also native speakers of Mandarin Chinese. They were not informed of the purpose of this study.

4.3.1. MOS evaluation

The performance of the overall quality samples was evaluated using the mean opinion score (MOS). Listeners were asked to rate the overall naturalness and prosodic appropriateness of samples on a scale from 1 and 5. Speech samples for the CSMSC test set were generated using both Tacotron 2 models. Two sentences in the test set were removed due to potentially inappropriate content. Then these synthesized samples were mixed with real speech samples and presented to listeners independently in random order.

4.3.2. Tone-3 sandhi

In this task, listeners were asked to judge whether there were Tone-3 sandhi errors in the speech samples while ignoring any acoustic artifacts not related to lexical tones. Specifically, for phrase level stimuli, they were also asked to identify the number of tonal errors in the samples.

5. Results

5.1. MOS

Table 2 summarizes the results of listeners' evaluation of the generative samples. As suggested in [25], paired Mann-Whitney U Tests were used to test pairwise comparisons for naturalness and prosody over three types of speech samples, and all comparisons were significant ($p \leq .001$) after applying the conservative Bonferroni correction for multiple comparisons. These results show that both models can synthesize highly natural speech samples with acceptable prosody and, similar to previous studies [19, 18], Tacotron 2 with BERT embeddings improved the naturalness and prosody of generated speech over the baseline Tacotron 2. It is also noticed that Tacotron 2 with BERT embeddings tends to make slightly more pronunciation errors than the baseline Tacotron 2.

Table 2: *Mean MOS of speech samples*

Models	Naturalness	Prosody
Tacotron 2	3.65	3.86
Tacotron 2+BERT	4.04	4.21
Ground Truth	4.39	4.53

5.2. Tonal coarticulation

As both models capture equally well the surface variations of lexical tones, only results from Tacotron 2 are reported for illustration. Regression models were fitted to each tone in each condition through locally estimated scatterplot smoothing (LOESS) and resulting curves with shaded 95% confidence intervals were presented in Fig. 3 and 4. There are strong (assimilatory) carry-over effects of the offset of the previous tone (Fig. 3), as the overall pitch is significantly lowered or raised by the low or high offset of the previous tone, respectively. Tone-3 + Tone-3 combinations were excluded in the analysis because it is a categorical tone sandhi process examined in next section. Fig. 4 indicates that anticipatory effects are weaker than the carry-over effects: the tone contours do not differ much before different onsets of the upcoming tone. Moreover, the anticipatory coarticulation is mostly dissimilatory, in which low onset f0 of the upcoming tone raises the f0 maximum of the previous tone. The findings closely mirror those obtained from native speakers [14] even in fine-grained details, suggesting that these coarticulatory patterns are learned and faithfully reproduced in TTS models in novel sentences.

[2] https://github.com/NVIDIA/tacotron2

[3] https://github.com/NVIDIA/waveglow

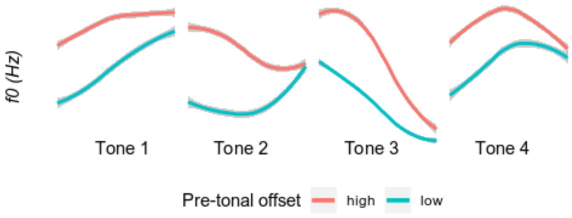

Figure 3: *Assimilatory carry-over coarticulation produced by Tacotron 2. Tone onsets are strongly affected by the offsets of previous tones.*

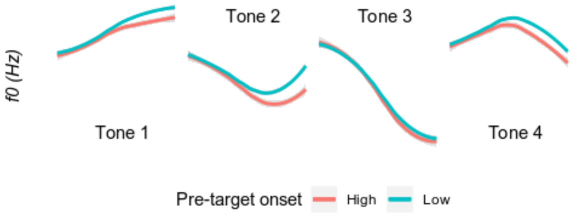

Figure 4: *Dissimilatory anticipatory coarticulation produced by Tacotron 2. Tone offsets are only slightly affected by the onsets of upcoming tones.*

5.3. Tone-3 sandhi

The Tone-3 sandhi experiment presents an interesting case concerning the phonological ability of TTS models. Mixed logistic regression models were fitted to test whether the difference between two models was significant. Both TTS models made few Tone-3 sandhi mistakes on bisyllabic words, even when such information is not specified in the Pinyin inputs. But Tacotron 2 is more accurate in synthesizing bisyllabic Tone-3 words than Tacotron 2 with BERT embeddings ($\hat{\beta} = 2.35, p \leq 0.001$).

The performance decreased for more complex stimuli, in which tone sandhi patterns interact with syntactic structures. Both models perform nearly equally poorly on trisyllabic words ($\hat{\beta} = -0.05, p = 0.82$), with mistakes unlikely to be made by native speakers. However, Tacotron 2 with BERT embeddings performs better than the phoneme-only Tacotron 2 on complex sentences ($\hat{\beta} = -0.52, p = 0.01$). On average, Tacotron 2 with BERT made 0.92 mistakes per phrase, while the baseline Tacotron 2 had 1.31 mistakes per phrase. This implies that at least some syntactic information encoded in BERT embeddings has been utilized in parsing long sentences, but the performance enhancement is limited with the current architecture.

Error analysis shows that the TTS models are not simply applying a fixed Tone-3 sandhi pattern to all trisyllabic words. For different $[\sigma_3\sigma_3\sigma_3]_w$ trigrams, both TTS models are capable of varying their predictions either as $[\sigma_3\sigma_2\sigma_3]_w$ or $[\sigma_2\sigma_2\sigma_3]_w$ depending on the actual structure of the phrases, evidence that the TTS models have learned to generalize the rules over novel instances rather than simply memorizing a fixed pattern. But the low accuracy suggests that such generalizability is limited. In long sentences, most tonal errors arise from changing tones at incorrect syntactic boundaries rather than not changing tones at all. Both TTS models consistently applied Tone-3 sandhi rules to every bigram or trigram in a sentence, but the errors are likely to be caused by the failure to parse a sentence into its correct hierarchical structure. These results suggest that Tacotron 2 models, though capable of generating highly natural speech (including prosodic contours), still could not be used as a model for categorical Tone-3 sandhi in Mandarin.

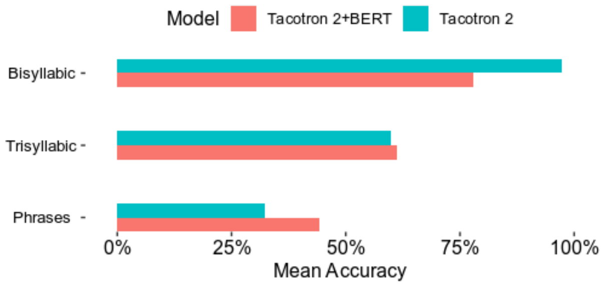

Figure 5: *Mean accuracy of Tone-3 sandhi*

6. Discussion

In this study, linguistically informed stimuli have been used to probe the phonetic and phonological knowledge of TTS models. While both TTS models can produce the surface tonal coarticulations that closely match results from human speakers [14], both models still fail to correctly apply the Tone-3 sandhi rules to unseen complex stimuli. Supplying syntactic information through BERT embeddings helps but the overall accuracy of Tone-3 sandhi in phrases was still low, suggesting that the model does not fully acquire the underlying linguistic rules in the current data-driven approach, as shown in some similar NLP studies [8, 7, 26]. The rules induced by neural networks might not be as straightforward as those clear-cut algebraic rules summarized in linguistics [27]. But this could also be caused by insufficient relevant instances in the training data, as the stimuli in this experiment are not highly frequent in daily conversations. These results also highlight the importance of developing linguistically informed stimuli to evaluate the performance of TTS models and of designing training data to cover as many phonological processes as possible.

In light of the results, TTS models seem to provide a complementary method to phonetic experiments for some research questions. The patterns of tonal coarticulation almost perfectly replicate what has been found in classic studies [14]. Arguably TTS models can be applied to mine linguistic knowledge hidden in large quantities of coarsely annotated text-speech pairs, which are relatively easier to acquire than time-aligned phonemic transcriptions. Compared to the representation learning approach [11, 12], the use of TTS models directly addresses the problem of plausibility, as researchers can verify the speech samples through listening. Given the limitations of current TTS models, the research questions that can be explored in this approach are also limited, but nonetheless, with proper evaluation, they can be applied to generate novel hypotheses or provide additional evidence to experimental results.

7. Acknowledgements

I would like to express my gratitude to Professor Patrice Beddor and Professor Jelena Krivokapić at University of Michigan for their helpful advice. I also thank my fellow linguists who had carefully listened to every generated sample in the experiment.

8. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2018, pp. 4779–4783.

[2] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality tts with transformer," *arXiv preprint arXiv:1809.08895*, 2018.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *arXiv preprint arXiv:1905.09263*, 2019.

[4] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[5] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[6] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 3617–3621.

[7] R. Futrell, E. Wilcox, T. Morita, and R. Levy, "Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency," *arXiv preprint arXiv:1809.01329*, 2018.

[8] K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni, "Colorless green recurrent networks dream hierarchically," *arXiv preprint arXiv:1803.11138*, 2018.

[9] G. Prasad, M. van Schijndel, and T. Linzen, "Using priming to uncover the organization of syntactic representations in neural language models," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 66–76.

[10] T. Linzen, "What can linguistics and deep learning contribute to each other? response to pater," *Language*, 2019.

[11] B. Li, J. Y. Xie, and F. Rudzicz, "Representation learning for discovering phonemic tone contours," *arXiv preprint arXiv:1910.08987*, 2019.

[12] B. Gerazov, G. Bailly, O. Mohammed, Y. Xu, and P. N. Garner, "A variational prosody model for mapping the context-sensitive variation of functional prosodic prototypes," *arXiv preprint arXiv:1806.08685*, 2018.

[13] Y. Chen and Y. Xu, "Production of weak elements in speech–evidence from f0 patterns of neutral tone in standard chinese," *Phonetica*, vol. 63, no. 1, pp. 47–75, 2006.

[14] Y. Xu, "Contextual tonal variations in mandarin," *Journal of phonetics*, vol. 25, no. 1, pp. 61–83, 1997.

[15] X. S. Shen, "Tonal coarticulation in mandarin," *Journal of Phonetics*, vol. 18, no. 2, pp. 281–295, 1990.

[16] Y.-F. Chien, J. A. Sereno, and J. Zhang, "Priming the representation of mandarin tone 3 sandhi words," *Language, Cognition and Neuroscience*, vol. 31, no. 2, pp. 179–189, 2016.

[17] C. Shih, "Mandarin third tone sandhi and prosodic structure," *Linguistic Models*, vol. 20, pp. 81–124, 1997.

[18] W. Fang, Y.-A. Chung, and J. Glass, "Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models," *arXiv preprint arXiv:1906.07307*, 2019.

[19] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu, "Pre-trained text embeddings for enhanced text-to-speech synthesis," *Proc. Interspeech 2019*, pp. 4430–4434, 2019.

[20] S. Tyagi, M. Nicolis, J. Rohnke, T. Drugman, and J. Lorenzo-Trueba, "Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection," *arXiv preprint arXiv:1912.00955*, 2019.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu, "Pre-training with whole word masking for chinese bert," *arXiv preprint arXiv:1906.08101*, 2019.

[23] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 215–10 224.

[24] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[25] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores." in *Interspeech*, 2017, pp. 3976–3980.

[26] R. Futrell, E. Wilcox, T. Morita, P. Qian, M. Ballesteros, and R. Levy, "Neural language models as psycholinguistic subjects: Representations of syntactic state," *arXiv preprint arXiv:1903.03260*, 2019.

[27] M. Baroni, "Linguistic generalization and compositionality in modern artificial neural networks," *arXiv preprint arXiv:1904.00157*, 2019.