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

Fine-grained prosody modeling in neural speech synthesis using ToBI representation

Yuxiang Zou, Shichao Liu, Xiang Yin, Haopeng Lin, Chunfeng Wang, Haoyu Zhang, Zejun Ma

Bytedance AI-Lab, China
{zouyuxiang.zyx, liushichao, yinxiang.stephen}@bytedance.com

Abstract

Benefiting from the great development of deep learning, modern neural text-to-speech (TTS) models can generate speech indistinguishable from natural speech. However, The generated utterances often keep an average prosodic style of the database instead of having rich prosodic variation. For pitch-stressed languages, such as English, accurate intonation and stress are important for conveying semantic information. In this work, we propose a fine-grained prosody modeling method in neural speech synthesis with ToBI (Tones and Break Indices) representation. The proposed system consists of a text frontend for ToBI prediction and a Tacotron-based TTS module for prosody modeling. By introducing the ToBI representation, we can control the system to synthesize speech with accurate intonation and stress at syllable level. Compared with the two baselines (Tacotron and unsupervised method), experiments show that our model can generate more natural speech with more accurate prosody, as well as effectively control the stress, intonation, and pause of the speech.

Index Terms: speech synthesis, ToBI representation, prosody modeling, intonation, stress and pause control

1. Introduction

Recently, neural text-to-speech (TTS) systems [1, 2, 3, 4] can generate speech that is indistinguishable from natural speech. However, the generated utterances are not expressive enough and the prosody may even be wrong in some situations, which may lead to misunderstanding. For example, for the sentence, ”You are going to the park”, it has a completely different meaning when reading it in a falling tone or a rising tone. The former expresses affirmation, and the latter expresses question and uncertainty. In many scenarios, generating utterances with correct and rich prosody is quite important. For dialogue interaction, rich prosody variation will make users feel cordial, natural, and not tired. For artificial intelligence (AI) education scenarios, both intonation and stress are required to be expressed accurately in order to convey correct information and knowledge to listeners.

From a linguistic point of view, prosody represents suprasegmentals of speech, including linguistic functions such as intonation, tone, stress, and rhythm. From an acoustic point of view, prosody is related to pitch, phone duration, energy, and spectral tilt [5]. Recent developments in end-to-end TTS systems have provided some unsupervised methods for prosody modeling. In terms of the granularity of prosody modeling, these methods can be divided into two categories: coarse-grained at sentence level [6, 7, 8] and fine-grained at word and phone levels [9, 10, 11]. Stanton et al. [7] proposed the Text-Predicted Global Style Token (TP-GST) architecture to predict style embedding from text which can control the speaking style at sentence level. However, the controlled attributes of the speech are ambiguous. In contrast, the method in [8] can control prosody conditioned on acoustic speech features which have clear physical attributes, such as pitch, pitch range, phone duration, energy, and spectral tilt. Similar to TP-GST, it still only provides sentence-level prosody control. Sun et al. [10] proposed a hierarchical variational auto-encoder (VAE) [9] structure for fine-grained prosody modeling which provided word-level and phone-level prosody control. However, it can neither predict appropriate prosody from text nor intuitively control intonation. Besides, some other neural models focus on prosody improving. CHiVE [12] uses linguistically driven dynamic hierarchical conditional VAE to produce varied prosody and other papers [13, 14] use BERT [15] derived features for prosody improving.

In addition to the unsupervised methods, linguists with rich expert knowledge have done a lot of research on prosody representation and developed ToBI (Tones and Break Indices) [16], a standard for English prosody labeling. Researchers explored the explicit use of ToBI in unit selection-based TTS, which enabled prosody modeling and generated expressive speech [17, 18, 19]. These approaches had a certain control ability of intonation, and the rate of successfully controlling the questioning tone in yes-no questions achieved 85% [19]. However, the naturalness of speech generated by these traditional methods is generally poor with a mean opinion score (MOS) of around 3.5.

To effectively utilize ToBI for prosody modeling as well as synthesizing speech with high naturalness, we propose a fine-grained prosody modeling system based on an end-to-end neural network. The proposed system consists of a ToBI prediction module and a Tacotron-based speech synthesis module [2]. ToBI prediction module uses a pretraining language model to predict ToBI labels, and the Tacotron model is used for prosody modeling. Our experiments show that our model can effectively control the pause, intonation, and stress of the speech, generating utterances with a richer prosodic variation. Also, compared with Tacotron baseline and the unsupervised method [8], the proposed method can significantly reduce prosody error rate, and generate more natural speech.

The main contributions of this paper are as follows:

- We use a pre-trained language model to predict ToBI labels, which could automatically predict prosody from text at word and syllable levels.
- Compared with Tacotron baseline and unsupervised method, our model can generate more natural speech with lower prosodic error rate.
- To our best known, this is the first end-to-end English speech synthesis system combined with ToBI representation.

The structure of this paper unfolds as follows. We first review ToBI in Section 2 and introduce the proposed method in

http://dx.doi.org/10.21437/Interspeech.2021-883

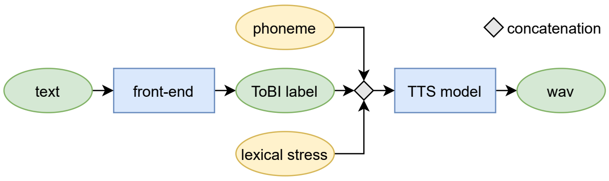

Figure 1: *The overall architecture of the proposed model.*

Section 3. Section 4 gives experimental comparison and results analysis, and Section 5 draws conclusions and future work.

2. Tones and Break Indices (ToBI)

ToBI (Tones and Break Indices) [16] is a set of conventions for transcribing and annotating the prosody of speech. Tones represent an analysis of the tonal events in terms of H and L, corresponding most closely to the utterance's intonation pattern. Tonal events include pitch accents, phrase accents, and boundary tones. Break indices mark the prosodic grouping of the words in an utterance and show the degree of disjuncture between adjacent words on a scale from 0 to 4. More details are described as follows:

- **Pitch accents** mark the stressed syllable of specific words that carry the most information in a sentence. Pitch accents include H* (high accent), L* (low accent), L*+H (a syllable which starts with a low accent and then rises) and L+H* (again low-high on one syllable, but with the second part accented).
- **Boundary tones** describe the pitch trend at each full intonation phrase boundary. The default is H% (high tone) and L% (low tone).
- **Phrase accents** describe the pitch movement between the ultimate pitch accent and the boundary tone. The default is H- and L-.
- **Break indices** represent the degree of disjuncture between adjacent words, examples of which are 4 for a full intonational phrase break, marked with L% or H%, at the end of a phrase or sentence, 3 for an intermediate phrase break, marked with H- or L-, and 1 for most phrase-internal word boundary.

Phrase accents and boundary tones together form phrasal tones which affect intonations of utterance, and pitch accents and break indices affect stress and pause of utterance respectively. Following the convention of ToBI, we transcribe and annotate the prosody of an internal TTS dataset. Therefore, the dataset includes not only text-speech pairs but also ToBI labels. The details of the dataset are described in Section 4.1.

3. Proposed approach

The proposed system consists of two components, as shown in Figure 1: (1) a ToBI prediction frontend which predicts ToBI tags from an input text sequence, and (2) a TTS model, which consists of a Tacotron-based acoustic model and a vocoder, generates waveform from input phoneme sequence with the predicted ToBI tags. In the training phase, the ToBI labels are used as the target of the frontend network as well as one of the input to the acoustic model. In the inference phase, the frontend converts the text sequence to ToBI tags, and then the phoneme sequence and the predicted ToBI tags are fed to the backend TTS model to generate speech.

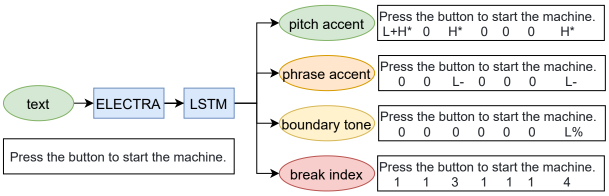
pitch accent
Press the button to start the machine.
L+H* 0 H* 0 0 0 H*
phrase accent
Press the button to start the machine.
0 0 L- 0 0 0 L-
text
ELECTRA
LSTM
boundary tone
Press the button to start the machine.
0 0 0 0 0 0 L%
Press the button to start the machine.
break index
Press the button to start the machine.
1 1 3 1 1 1 4

Figure 2: *Model structure of ToBI prediction frontend.*

3.1. ToBI prediction frontend

ToBI prediction frontend aims to automatically predict the appropriate prosody from the text, including pause, stress, and intonation. Different from automatic ToBI annotation [20] which detects and classifies the prosodic events of the ToBI standard from speech, our module predicts ToBI tags only from text. Since ToBI labels are word-level tags, ToBI prediction task is actually a sequence tagging task.

Recently, unsupervised representation learning models have shined in the field of natural language processing (NLP) and achieved state-of-the-art results in downstream tasks [15, 21, 22]. Due to the small amount of training data and the complexity of ToBI features, a self-supervised language representation model, ELECTRA [22], is chosen for ToBI prediction. ELECTRA model is trained to distinguish ”real” input tokens vs ”fake” input tokens generated by another neural network, similar to the discriminator of a Generative Adversarial Nets (GAN) [23]. Using the pre-trained model ELECTRA is conducive to extracting the grammatical and semantic information in the text, which is helpful for the prediction of ToBI.

Figure 2 shows the model structure of ToBI prediction frontend, including four prediction tasks of pitch accent, phrase accent, boundary tone and break index. We use the ToBI-labeled text to finetune ELECTRA model on WordPiece level. The input of the ELECTRA is WordPiece token represented by a 768-dimensional embedding. The output of the ELECTRA is followed by a LSTM as the tagging model. The loss function is cross-entropy loss between the ground truth and prediction of the ToBI sequence.

3.2. Speech synthesis backend

The baseline neural speech synthesis model is Tacotron [2], an end-to-end architecture consisting of encoder and decoder with attention mechanism. The encoder maps an input text sequence to a series of hidden representations consumed by the decoder to predict a spectrogram sequence. Finally, a neural vocoder is employed to synthesize the waveform from the predicted spectrogram.

Compared with the original Tacotron, we also use ToBI labels as a part of the input features to the encoder, considering that encoder should be able to capture text-related prosodic information. Overall, the input features consist of phoneme, lexical stress, break index, pitch accent, phrase accent, and boundary tone. Figure 3 shows the input features and the corresponding expanding mechanism in speech synthesis backend. Specifically, lexical stress is marked on vowels, where 0 means no stress, 1 means primary stress, and 2 means secondary stress. Break index is marked on the last phoneme of the word. Pitch accent is marked on the primary stress syllable and then expanded to all phonemes in that syllable. Phrase accent and boundary tone are marked on the last word of the phrase boundary and then expanded to all phonemes in that word. These six

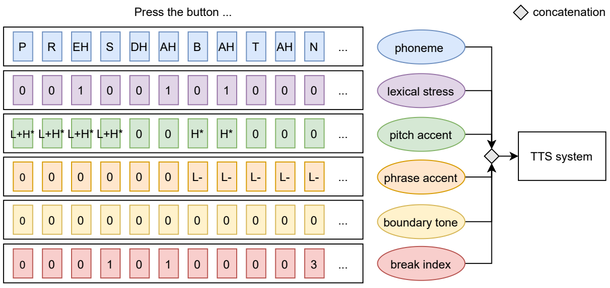

Figure 3: *The input features and the expanding mechanism of speech synthesis backend.*

features are first embedded into vectors respectively and then concatenated to form the input of the encoder. By introducing ToBI features into training, we can control the system to synthesize speech with different prosodies by giving different ToBI labels during inference. Since the ToBI features act on word and syllable levels, they can play a fine-grained control on the synthesized speech.

4. Experiments and results

4.1. Experimental Setup

An internal high-quality English dataset recorded by a professional male native speaker is used in our experiments. This dataset consists of about 15000 sentences with a sampling rate of 24khz, for a total duration of 17 hours. The dataset is annotated by 19 people trained by linguists on the aspect of ToBI.

For ToBI prediction frontend, we train our frontend model by finetuning ToBI tagging task on the pre-trained ELECTRA-BASE model. This model consists of 12 layers with a hidden size of 768 and the model parameter size is 110M. We follow fine-tuned hyperparameters described in [22]. To evaluate the model, we divide the dataset into a training set of 13000 sentences and a validation set of 2000 sentences.

For acoustic model, the target acoustic features are log magnitude spectrogram extracted with 50ms frame length, 12.5ms frame shift. The inputs of the acoustic model are composed of the labels of phoneme, lexical stress, break index, pitch accent, phrase accent, and boundary tone. These features are embedded into 448, 64, 32, 32, 32, and 32 dimensions respectively, and then concatenated together as the encoder input. We evaluate the acoustic model on an out-of-domain test set consisting of 200 sentences.

Four end-to-end speech synthesis systems are built for comparison:

- **TACO**: the model architecture is the same with Tacotron [2] and the input features are simply composed of the labels of phoneme, lexical stress, and word boundary.
- **TP-DPE**: this model is based on [8] with some modifications by extending sentence-level prosody control to phone-level. Specifically, phone-level prosodic features, including phone duration, pitch, and energy (DPE features), are used to realize prosody modeling. While training, ground-truth DPE features are fed to decoder as additional features, and encoder is trying to predict those features. While synthesizing, the predicted features are fed to decoder directly. This method is called text-predict DPE (TP-DPE for short).
- **TP-ToBI**: the proposed method introduced in Section 3. The input features are composed of phoneme, lexical stress, break index, pitch accent, phrase accent, and boundary tone, where the ToBI-related labels are predicted from text using the proposed ToBI prediction frontend.
- **GT-ToBI**: the acoustic model and the input feature are the same as TP-ToBI, but the ToBI-related labels are manually revised from the results of ToBI prediction frontend.

We train all the acoustic models for 2 million steps using a single GPU with a batch size of 64. Finally, a WaveRNN [24] neural vocoder is used to synthesize waveform from the mel spectrograms in real-time and high fidelity.

4.2. Results

In this part, we evaluate the proposed method in three aspects. First, ToBI prediction frontend is evaluated. Next, we analyze the controllability of the speech synthesis backend, and finally, we compare the performance of the cascaded system with other baseline systems. [1].

4.2.1. Evaluation of ToBI prediction frontend

We evaluate the performance of the four prediction tasks on the test set from two perspectives, including objective and subjective evaluation. Objective evaluation calculates the accuracy and F1 score using ground truth labels. For subjective evaluation, we try to evaluate human acceptability for the prediction. Specifically, some of the predicted ToBI tags are modified to suitable ones judged by humans based on the principle of the minimum number of revisions. Then, the accuracy and F1 score are calculated between the predicted tags and the corresponding revised version. All metrics are calculated at word level.

The evaluation results demonstrated in Table 1 show that our model has a high prediction accuracy. It is worth noting that although objective assessments on macro F1 are low, the corresponding subjective results are high. This is because humans may read the same text with different prosodies and the ground truth may only represent one type of them. It proves the predicted results are highly acceptable by humans. We can also find that accuracy is much higher than macro F1. This is because the proportion of different categories in the dataset is extremely unbalanced. Take pitch accent as an example, the proportion of H* in the dataset is 33.9%, but the proportion of L* is 2.7%.

Table 1: *Evaluation of ToBI prediction frontend. Accuracy and F1 are calculated at word level. X/Y means that the former is objective evaluation calculated on the ground truth labels and the latter is subjective evaluation reflecting human acceptability. (Accuracy equals micro F1.)*

ToBI	**accuracy**	**macro F1**
pitch accent	0.86 / 0.99	0.54 / 0.96
phrase accent	0.95 / 0.96	0.77 / 0.87
boundary tone	0.98 / 0.99	0.77 / 0.97
break index	0.94 / 0.94	0.76 / 0.82

[1] Audio samples available on https://sysuzyx.github.io/TTSwithTOBI

4.2.2. Controllability evaluation of speech synthesis backend

In this part, we individually test the controllability of the backend speech synthesis model, that is, under the given ground truth ToBI labels, test whether each ToBI feature is effectively reflected in generated speech. In the controllability test, we use 200 out-of-domain sentences with ToBI labels to synthesize speech for evaluation. Then, by comparing generated speech with ground truth, listeners professed in linguistic judge whether each pitch accent, phrasal tone and break index is naturally realized based on their own auditory impression. For each test utterance, unless every target ToBI feature is naturally realized and others are not, we will not think this utterance succeeds on ToBI prosody control.

Table 2 shows the success rate of sentence-level prosody control. The success rate of pitch accent, phrasal tone and break index is 87%, 93%, and 97.5% respectively. And for pitch accent, it is slightly lower than the other two. According to listeners' feedback, some of the words without labeled H* are pronounced with a high accented sense of hearing. However, most of the generated speech is still acceptable in naturalness with these extra high accents. The experiment results indicate that the introduction of ToBI features in the acoustic model (concatenated at the input of the encoder) is simple but effective.

Table 2: *The sentence-level prosody control success rate of the backend speech synthesis system. Phrasal tone consists of phrase accent and boundary tone as described in Section 2.*

	pitch accent	**phrasal tone**	**break index**
success rate	87%	93%	97.5%

4.2.3. Subjective evaluations of different systems

We perform AB preference tests in terms of naturalness to evaluate the performances of the 4 different systems described in Section 4.1. 10 listeners proficient in English participated in the evaluation using headphones. Each listener evaluated 40 pairs of utterances synthesized from the two comparative systems. After listening to each pair of synthesized utterances, the listeners were asked to choose their preferred one. They could choose "N/P" if they had no preference.

Table 3 shows the AB preference results of different systems on naturalness. The results show that the proposed model significantly outperforms Tacotron baseline and unsupervised TP-DPE method. For TP-ToBI and GT-ToBI systems, there is no statistically significant preference (p>0.05). This indicates that the predicted ToBI labels and the ground truth are equally acceptable. AB test reflects the overall preference between the two systems and a delicate evaluation and analysis on prosodic performance will be introduced in the next section.

Table 3: *AB preference scores(%) of different systems on naturalness. "N/P" stands for no preference. p denotes the p-value of a t-test between two systems.*

TACO	TP-DPE	TP-ToBI	GT-ToBI	N/P	p
16.25%	-	**53.75%**	-	30%	<e-3
-	6%	**77.25%**	-	16.75%	<e-3
-	-	22.75%	13%	64.25%	0.075

Table 4: *The sentence-level error rate of stress, intonation, and pause.*

model	**stress**	**intonation**	**pause**
TACO	20%	16%	11%
TP-DPE	28%	18%	6%
TP-ToBI	6%	5%	6%
GT-ToBI	1%	1%	1%

4.2.4. Error feedback evaluations on prosody

In order to compare the different systems for more detail, we also conduct an error feedback evaluation on prosody. In error feedback evaluation, listeners give specific feedback on the three types of errors: stress, intonation, and pause, which are related to pitch accent, phrasal tone, and break index respectively. Specifically, stress error means that the stressed word is not stressed, or the word that should not be stressed is stressed. Intonation error means that the intonation is wrong or unnatural, such as a yes-no question without a rising tone or an exclamation without a falling tone. Pause error means that there is no pause or the pause is too short or too long where the pause should be.

The test set is composed of 200 out-of-domain sentences, including 50 wh-questions, 50 yes-no questions, 50 exclamations, and 50 declarations. The lengths of the sentences range from 3 to 25. The possible prosodic errors described above were marked by listeners when listening speech synthesized by the four systems, and the error rates were calculated according to their feedback. The error rate is the proportion of the number of sentences with the defined error to the total number of sentences in the test set.

Table 4 shows the error rate of stress, intonation, and pause of the four systems. As shown in Table 4, the proposed TP-ToBI model can substantially reduce prosodic errors compared with Tacotron baseline and unsupervised TP-DPE method. Besides, the results shown in the last row of Table 4 demonstrate that by revising the predicted ToBI labels, the error rate of the stress, intonation, and pause can be reduced to almost 0. This is of great significance for correcting bad cases in actual application scenarios.

5. Conclusion and future work

In this paper, we explore a new method to enhance naturalness and controllability of E2E-TTS system with ToBI features. We propose a fine-grained prediction and control method in neural speech synthesis with ToBI representation. The proposed system consists of a text frontend for ToBI prediction and a Tacotron based TTS module for prosody modeling. Experiments show that our model can effectively control the stress, intonation, and pause of the speech. Also, compared with the unsupervised method, it can synthesize more natural speech and significantly reduce the error rate of prosody. In future work, we will explore automatic ToBI annotation to alleviate the manual labeling work. Besides, we will further explore other applications, such as multi-speaker prosody modelling.

6. Acknowledgements

The authors would like to thank Xiao Mei and Ying Liang for providing expert knowledge in linguistics and data annotations and evaluation.

7. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[2] 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.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[4] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling," *arXiv preprint arXiv:2010.04301*, 2020.

[5] C. A. De, D. L. Bolinger, D. Gibbon, E. Garding, J. T'Hart, N. Gronnum, S. Alcoba, J. Murillo *et al.*, *Intonation systems: a survey of twenty languages*. Cambridge University Press, 1998.

[6] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[7] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.

[8] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *Proc. Interspeech 2020*, pp. 4432–4436, 2020.

[9] W.-N. Hsu, Y. Zhang, R. Weiss, H. Zen, Y. Wu, Y. Cao, and Y. Wang, "Hierarchical generative modeling for controllable speech synthesis," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rygkk305YQ

[10] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6264–6268.

[11] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5911–5915.

[12] T. Kenter, V. Wan, C.-a. Chan, R. Clark, and J. Vit, "Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *ICML*, 2019.

[13] T. Kenter, M. Sharma, and R. Clark, "Improving the prosody of rnn-based english text-to-speech synthesis by incorporating a bert model," *Proc. Interspeech 2020*, pp. 4412–4416, 2020.

[14] Y. Xiao, L. He, H. Ming, and F. K. Soong, "Improving prosody with linguistic and bert derived features in multi-speaker based mandarin chinese neural tts," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6704–6708.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[16] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Second international conference on spoken language processing*, 1992.

[17] C. W. Wightman, A. K. Syrdal, G. Stemmer, A. Conkie, and M. Beutnagel, "Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis," in *Sixth International Conference on Spoken Language Processing*, 2000.

[18] J. F. Pitrelli and E. M. Eide, "Expressive speech synthesis using american english tobi: questions and contrastive emphasis," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 694–699.

[19] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The ibm expressive text-to-speech synthesis system for american english," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1099–1108, 2006.

[20] A. Rosenberg, "Autobi-a tool for automatic tobi annotation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[22] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *ICLR*, 2020. [Online]. Available: https://openreview.net/pdf?id=r1xMH1BtvB

[23] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[24] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.