Interspeech 2024
1-5 September 2024, Kos, Greece

Multi-Modal Automatic Prosody Annotation with Contrastive Pretraining of Speech-Silence and Word-Punctuation

Jinzuomu Zhong[1,2], *Yang Li*[2], *Hui Huang*[2], *Korin Richmond*[1], *Jie Liu*[2], *Zhiba Su*[2], *Jing Guo*[2], *Benlai Tang*[2], *Fengjie Zhu*[2]

[1]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]Department of AI Technology, Transsion, China

j.zhong-12@sms.ed.ac.uk, korin.richmond@ed.ac.uk, benlai.tang@transsion.com

Abstract

In expressive and controllable Text-to-Speech (TTS), explicit prosodic features significantly improve the naturalness and controllability of synthesised speech. However, manual prosody annotation is labor-intensive and inconsistent. To address this issue, a two-stage automatic annotation pipeline is novelly proposed in this paper. In the first stage, we use contrastive pretraining of Speech-Silence and Word-Punctuation (SSWP) pairs to enhance prosodic information in latent representations. In the second stage, we build a multi-modal prosody annotator, comprising pretrained encoders, a text-speech fusing scheme, and a sequence classifier. Experiments on English prosodic boundaries demonstrate that our method achieves state-of-the-art (SOTA) performance with 0.72 and 0.93 f1 score for Prosodic Word and Prosodic Phrase boundary respectively, while bearing remarkable robustness to data scarcity.

Index Terms: Prosody Annotation, Contrastive Learning, Controllable Speech Synthesis

1. Introduction

Recent advances in expressive Text-to-Speech (TTS) systems have made it possible to generate speech that is indistinguishable from human speech [1, 2]. High naturalness of these systems is largely achieved by implicit modelling of prosody: 1) enriching text/phonetic representation with contextual information from text/phonetic pretraining such as [2, 3, 4, 5, 6]; 2) self-supervised learning of prosody-related acoustic features such as pitch or other learned information from audio [1, 7, 8, 9, 10]; 3) enhancing model capabilities with generative modelling [11, 12]. However, implict modelling of prosody suffers controllability issues as it cannot correct prosody and pause errors when the model generates bad cases.

By contrast, conditioning the speech on explicit prosodic features such as hierarchical prosodic boundaries [13, 14], ToBI [15, 16], and contrastive focus [17, 18] explicitly models speech prosody, and thus offer two advantages: 1) improvement in naturalness by fine-grained prosodic information, and more importantly, 2) precise control over prosody and pause as demonstrated in these previous works. However, human annotation of prosodic features is time-consuming, expensive, and often inconsistent [13] and therefore calls for automatic annotation.

Three categories of method are attempted to address automatic prosody annotation: 1) audio-only spectrogram analysis approach [15, 19], 2) text-only prosody prediction [16, 17, 20, 21], and 3) multi-modal prosody annotation [13, 14, 22, 23, 24]. Among all these, multi-modal prosody annotation with pretrained text-speech model outperforms the others by utilizing inputs of both modalities, but the result is still far from satisfactory. Dai et al. [13] use cross-attention to align and fuse acoustic information from a pretrained Conformer-ASR model [25] and text information from a pretrained BERT [26] for Chinese prosodic boundary annotation. Yuan et al. [14] apply cross-lingual transfer learning of the same model for Mongolian. The major drawbacks of this approach are two-fold. 1) Implicit alignment between text and speech by cross-attention leads to high requirement for annotated data. 2) Phonetic posteriorgrams (PPGs) lack sufficient prosodic information [27] which is crucial for prosody annotation. These drawbacks motivate us to search for better representation with enriched prosody information that can improve the performance of prosody annotation. In this paper, we study the automatic annotation of English prosodic boundaries and its effect on synthesised speech.

Inspired by cross-modal contrastive learning, CLIP [28], CLAP [29, 30], and its adaptation in the TTS community such as CLAPSpeech [9], we propose a two-stage training pipeline with novel pretraining strategy and simple model architecture that achieves SOTA performance on automatic annotation of English prosodic boundaries. In the first stage, we propose the contrastive text-speech learning of a new unit, named Speech-Silence and Word-Punctuation (SSWP) pairs, as silence- and punctuation-related information is crucial for detecting prosodic boundary. Conformer [31] and pretrained BERT [26] models are used as speech and text encoder respectively to extract the shared representation. In the second stage, i.e. prosody annotation, we add the text and audio latent representation of each SSWP and concatenate text-speech latent representations of all SSWPs in a sentence to represent speech prosody. Both modalities of inputs are used due to the variability of speech prosody given the same text. Finally, a standard bi-LSTM sequence classification network maps the combined text-speech latent representations to their corresponding prosodic boundary classes.

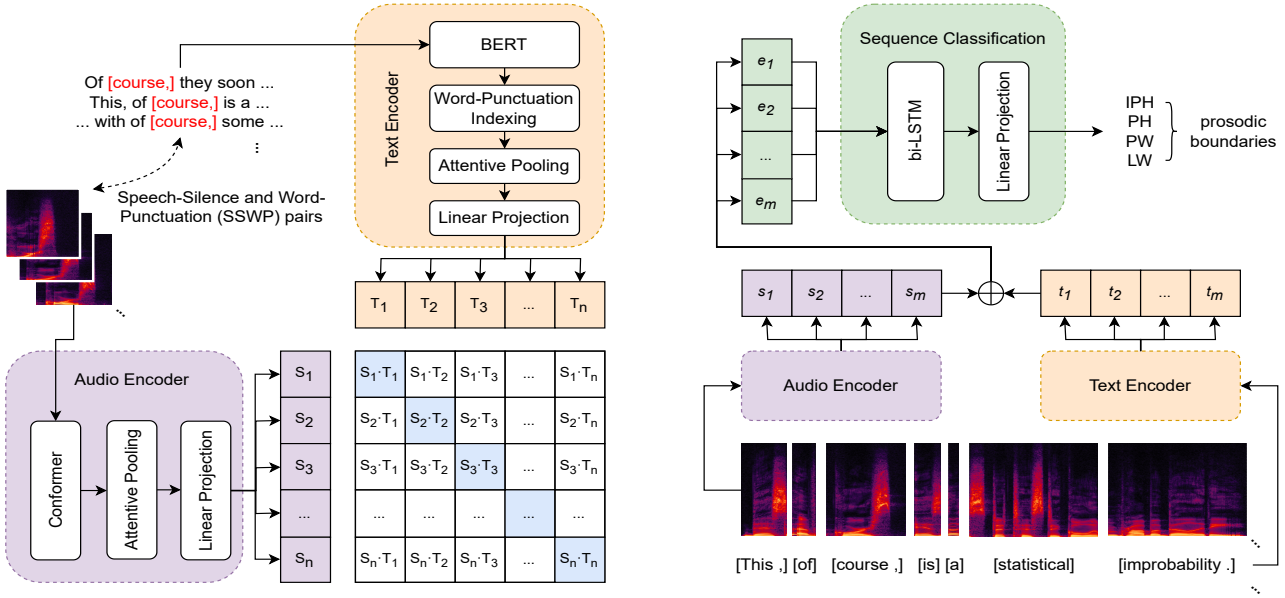
To summarize, our approach mainly contributes to the following three areas:

- To the best of our knowledge, we are the first to introduce contrastive learning to prosody annotation. We propose a novel pretraining of Speech-Silence and Word-Punctuation (SSWP) pairs to enhance the prosodic information of the learned representation.
- We propose a novel multi-modal prosody annotation architecture in the second stage, comprising pretrained encoders, straightforward yet effective text-speech feature fusion, and sequence classification network.
- We achieve SOTA performance on English posodic boundary annotation task with higher resilience against limited data.

2. Method

As shown in Figure 1, our proposed method automatically annotates prosodic boundaries via a two-stage training pipeline.

10.21437/Interspeech.2024-2133



(a) Stage I: Contrastive Text-Speech Pretraining of Speech-Silence and Word-Punctuation (SSWP) Pairs

(b) Stage II: Multi-Modal Prosody Annotation (Audio Encoder and Text Encoder are initialized from Stage I and finetuned in Stage II.)

Figure 1: *The architecture of our proposed two-stage training pipeline.*

In Section 2.1, we introduce the standard of English prosodic boundaries adopted in this work and how these features are annotated by human, a process that our approach aim to mimick. In Sections 2.2 and 2.3, we introduce the first stage contrastive text-speech pretraining of SSWP pairs and the second stage multi-modal prosody annotation, respectively. Code is openly available at `https://github.com/jzmzhong/Automatic-Prosody-Annotator-with-SSWP-CLAP/`

2.1. Prosodic Boundary Annotation

The prosody annotation adopted in this work categorizes the prosodic boundaries of English speech into four levels, including Lexicon Word (LW), Prosodic Word (PW), Prosodic Phrase (PPH), and Intonational Phrase (IPH) [32], as shown in Table 1. Human annotators label the boundaries based on both text (e.g. syntax, semantics) and speech (e.g. pitch, pace, silence).

Table 1: *Four-level prosodic boundaries of English speech.*

IPH	We must urge representatives to push for reforms.							
PPH	We must urge representatives				to push for reforms.			
PW	We must		urge representatives		to push for reforms.			
LW	We	must	urge	representatives	to	push	for	reforms .

2.2. Contrastive Text-Speech Pretraining of SSWP Pairs

The overall architecture of this first stage is shown in Figure 1(a). The pretraining aims to extract the prosodic space by contrasting text-speech pairs of the same phonetic content under different contexts.

Speech-Silence & Word-Punctuation (SSWP) Pairs Prosodic boundaries have a high correlation with both punctuations in text and silence segments in speech as they usually indicate the possible existence of PW, PPH, or IPH boundaries [33]. Neither prompts/sentences [29, 30] nor words [13] are ideal contrastive units for pretraining as they do not include boundary-related information. We propose to contrastively pretrain Speech-Silence & Word-Punctuation (SSWP) pairs, where words with following punctuations are paired with speech segments with following silence segments if there are any.

Text & Audio Encoder We choose Conformer [25] as the audio encoder and BERT [26] as the text encoder due to its wide application in various speech and text tasks, similar to the encoder choices made in [9, 29, 30]. Speech-silence segments are passed into Conformer blocks. An attentive pooling layer is used to handle variable length and a linear projection layer is used to reshape latent representations into the dimension of the joint prosodic space, following [9, 28, 29, 30]. The entire text utterances of text-punctuation segments are passed into BERT to include contextual information. A following indexing layer then selects the corresponding subwords for the SSWP in a sentence, following [9]. Similar to the audio encoder, an attentive pooling layer and a linear projection layer map the text representation to the joint prosodic space. Let the speech-silence segment of index i be $X_{SS}^{(i)}$, the text utterance that contains the corresponding word-punctuation be $X_{WP}^{(i)}$, and the subword start & end indexes of the word-punctuation in the utterance be (i_m, i_n). The text and speech representation of each SSWP, T_i and S_i can be written as follows.

$$T_i = \text{Linear}(\text{Pool}(\text{BERT}(X_{WP}^{(i)})[i_m : i_n])) \quad (1)$$

$$S_i = \text{Linear}(\text{Pool}(\text{Conformer}(X_{SS}^{(i)}))) \quad (2)$$

Contrastive Pretraining The text-speech model is trained with the contrastive learning paradigm between the text and speech embeddings of paired extended-words, T_i and S_i, following the same loss function as in [9, 28, 29, 30]:

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{2n}\sum_{i=1}^{n}(\log \frac{exp(S_i \cdot T_i/\tau)}{\sum_{j=1}^{n} exp(S_i \cdot T_j/\tau)} + \log \frac{exp(T_i \cdot S_i/\tau)}{\sum_{j=1}^{n} exp(T_i \cdot S_j/\tau)}) \quad (3)$$

where n is the number of samples in a batch, and τ is a learnable temperature parameter for scaling the loss. The negative examples T_j, S_j where $j \neq i$, are sampled randomly from a group of SSWP pairs with the same phonetic contents but different contexts.

2.3. Multi-modal Prosody Annotation

Since the same text can be expressed in different speech prosody dependant on the speaker, multi-modal information is used. The overall architecture of this second stage is shown in Figure 1(b). SSWP pairs in a sentence are passed into the Text & Audio Encoder pretrained from the first stage. The text and speech representations of a sentence, $t_{1,2,...m}$ and $s_{1,2,...m}$ are then simply added to obtain a latent representation with enhanced prosodic information $e_{1,2,...m}$. A standard bi-LSTM followed by a linear projection layer with softmax function is used to form the sequence classification network that predicts the probability distribution of prosodic boundaries $\hat{y}_{1,2,...m}$, written as follows:

$$\hat{y}_{1,2,...m} = \sigma(\text{Linear}(\text{biLSTM}(t_{1,2,...m} + s_{1,2,...m}))) \quad (4)$$

where σ denotes softmax activation. We adopt Cross Entropy Loss as the training objective of the classification model.

3. Experiments

3.1. Datasets

Pretraining An open-source English ASR corpus, LibriSpeech [34], is used for the first stage pretraining. 976.6 hours containing all subsets except for `test-clean` are used for training, and the remaining 5.4 hours are used for validation. Punctuation is inserted by using Whisper [35]. Word boundaries are obtained using Montreal Forced Aligner [36].

Prosody Annotation A proprietary English TTS corpus with prosodic boundary annotation is used for the second stage downstream prosody annotation due to the lack of high-quality English open-source corpus with prosodic boundary annotation. Performance of different systems in three scenarios are evaluated: on low/medium resource and on unseen/seen speakers, listed in Table 2. In the unseen speaker scenario, the trained model is used to predict prosodic boundaries of out-of-domain speakers; whereas in the seen speaker scenario, the trained model is used to predict prosodic boundaries of out-of-domain utterances by an in-domain speaker. Low resource unseen speaker is not evaluated since: i) zero-shot unseen speaker with limited one-speaker data is highly contingent upon the similarity in prosody between the unseen speaker and the trained speaker, ii) such a scenario is hardly encountered in the development of TTS systems. To validate/test the seen speaker scenario, we must select out-of-domain utterances from the same in-domain speaker corpus for training; due to limited data, an unavoidably small valid/test data size of 500 utterances is adopted.

Table 2: *Data composition of three annotation scenarios. Res. - Resource; Tar. - Target; Spk. - Speaker. Both the number of utterances and the number of speakers are listed, e.g. 21k (7spk) denotes 21,000 utterances by 7 speakers.*

Res. Scale	Tar. Spk.	Train	Valid	Test
Medium	Unseen	21k (7Spk)	2k (1Spk)	2k (1Spk)
Medium	Seen	21k (7Spk)	500 (1Spk)	500 (1Spk)
Low	Seen	8k (1Spk)	500 (1Spk)	500 (1Spk)

3.2. Baselines

We adopt two baselines: text-only prediction baseline named `Text Baseline`, and multi-modal annotation baseline named `Multi-modal Baseline`. For the `Text Baseline`, we use only text as input to a pretrained BERT[1] followed by a fully-connected layer as the classifier. For the `Multi-modal Baseline`, we reproduce the work in [13] as faithfully as possible and reapply the model on the same English dataset - both text and speech are passed into respective encoder and then combined using attention followed by a Transformer decoder and a fully-connected layer as classifier. We use the same aforementioned BERT as text encoder and a pretrained Conformer-ASR model[2] as audio encoder. To handle the mismatch in length between the latent representation of subword units in BERT and the prosodic labels of word units in classification targets, we adopt average pooling to merge the subword latent representations of each word. This is the only difference we have made in reproducing `Text Baseline` and `Multi-modal Baseline` and we do not think such adaption due to different experiment language would be an issue. We achieve comparable baseline results to the original paper, as shown in Table 3.

3.3. Training Configurations

Model Configurations The Conformer and the pretrained BERT that we adopt for the audio and text encoders of the `Proposed Work` are much smaller in size compared to those of the two baselines mentioned above. To save computing resources and speed up training, we use fewer parameters in our `Proposed Work` but still use openly available large pretrained models as baselines for fair and benevolent comparison. We use 12-layer BERT of 768 dimension size[1] and 18-layer Conformer of 512 dimension size[2] for the two baselines. We use 4-layer BERT of 256 dimension size[3] and 4 layers Conformer of 256 dimension size for our proposed work.

Pretraining We pretrain the text-speech model on 4 Nvidia 3090Ti GPUs with a batch size of 2,048 text-speech pairs (512 pairs per GPU). We use the Adam optimizer with an initial learning rate of 1e-4. We train the text-speech model for 50 epochs with a cosine learning rate schedule.

Prosody Annotation We train the classification model on 1 Nvidia 3090Ti GPU with a batch size of 16 sentences. We use the Adam optimizer with an initial learning rate of 1e-5. We train the classification model for 50 epochs with a cosine learning rate schedule.

4. Results

4.1. Objective Evaluation

The results of our proposed work, compared with previous benchmarks, are shown in Table 3. Due to limited space, the results of IPH boundary are not listed since they are all 0.99-1.00 in terms of f1 score. The results of LW boundary is not listed either as it is the default tag with an f1 score of 0.97-0.98. Different prosodic boundary annotators mainly differ in PW and PPH boundaries detection.

Our `Proposed Work` improves the `Multi-modal Baseline`, by **0.20-0.29** and **0.11-0.18** f1 score of PW and PPH boundaries respectively across all three annotation scenarios. Additionally, our proposed work achieves near-human expert annotation accuracy on the f1 score of PPH boundaries, ranging between 0.91-0.93. These results demonstrate both higher performance and broader applicability of our proposed work. Moreover, our proposed work has higher resilience to data scarcity. In low resource seen speaker scenario, the efficacy of the multi-modal baseline vanishes compared with text

[1] https://huggingface.co/bert-base-uncased

[2] https://huggingface.co/nvidia/stt_en_conformer_ctc_large

[3] https://huggingface.co/prajjwal1/bert-mini

Table 3: *Results of previous benchmarks and the proposed work. Res. - Resource, Spk. - Speaker, prec - precision, rec - recall.*

Systems	Medium Res. Unseen Spk.						Medium Res. Seen Spk.						Low Res. Seen Spk.					
	PW			PPH			PW			PPH			PW			PPH		
	prec	rec	f1	prec	rec	f1	prec	rec	f1	prec	rec	f1	prec	rec	f1	prec	rec	f1
Text Baseline	0.35	0.56	0.43	0.88	0.73	0.80	0.38	0.46	0.42	0.88	0.63	0.74	0.35	0.59	0.44	0.87	0.65	0.75
Multi-modal Baseline	0.44	0.48	0.46	0.84	0.83	0.84	0.50	0.43	0.46	0.86	0.75	0.80	0.35	0.59	0.44	0.87	0.63	0.73
Proposed Work	**0.76**	**0.58**	**0.66**	**0.94**	**0.93**	**0.93**	**0.70**	**0.74**	**0.72**	**0.91**	**0.93**	**0.92**	**0.70**	**0.75**	**0.73**	**0.93**	**0.89**	**0.91**

Table 4: *Results of ablation studies. Only one component is ablated at each time to better investigate its efficacy.*

Systems	Medium Res. Unseen Spk.						Medium Res. Seen Spk.						Low Res. Seen Spk.					
	PW			PPH			PW			PPH			PW			PPH		
	prec	rec	f1	prec	rec	f1	prec	rec	f1	prec	rec	f1	prec	rec	f1	prec	rec	f1
Proposed Work	**0.76**	0.58	**0.66**	**0.94**	0.93	**0.93**	**0.70**	**0.74**	**0.72**	**0.91**	0.93	**0.92**	**0.70**	**0.75**	**0.72**	0.93	0.89	0.91
w/o Contrastive Pretrain	0.74	0.48	0.58	0.89	**0.95**	0.92	0.68	0.68	0.68	0.88	0.93	0.91	0.66	**0.75**	0.70	**0.94**	0.89	**0.92**
w/o Any Pretrain	0.70	0.44	0.54	0.88	**0.95**	0.91	0.64	**0.74**	0.69	0.89	0.94	**0.92**	0.64	0.73	0.68	0.91	**0.90**	0.91
w/o SSWP	0.64	0.48	0.55	0.87	0.84	0.85	0.64	0.69	0.66	**0.91**	0.74	0.81	0.57	0.65	0.61	0.86	0.75	0.80
w/o bi-LSTM	0.65	**0.60**	0.62	0.92	0.91	0.91	**0.70**	0.58	0.64	**0.91**	**0.95**	0.88	0.55	0.61	0.58	0.93	0.80	0.86

baseline, while our proposed work can achieve remarkable 0.29 and 0.18 f1 score gain for PW and PPH boundaries respectively. Objective results prove that our proposed work, despite being smaller in model size, gives better performance, broader applicability, and higher resilience to data scarcity.

4.2. Ablation Studies

We conduct extensive ablation studies to show the efficacy of each module in our proposed work, as shown in Table 4. The best system `Proposed Work`, as shown on the first row, outperforms all ablated systems on most objective evaluation metrics, proving the necessity of each aspect of our proposed approach. `w/o Contrastive Pretrain` system does not use the pretrained text-speech model from the first stage to initialize classification model weights in the second stage, but the text encoder is still initialized by weights from the pretrained BERT. `w/o Any Pretrain` system does not use any pretrained model to initialize classification model weights. `w/o SSWP` system uses tokens instead of SSWP as text-speech pairs both in the first stage and second stage. `w/o bi-LSTM` system removes the bi-LSTM network from the best system and relies on one linear projection layer to classify prosodic boundaries.

Contrastive pretraining, SSWP, and sequence classification are found to be significantly contributing factors, as evidenced by the comparison between the best system and the ablated systems:

1) Contrastive pretraining contributes to a 0.08 increase, compared with BERT pretraining alone, and a 0.12 increase, compared with no pretraining, in PW boundary f1 score under the medium resource unseen speaker scenario. This is due to the richer prosody representation extracted from contrastive pretraining in the first stage, compared with pure BERT or unpretrained representations. Under the medium resource seen speaker scenario, the effect of contrastive pretraining drops to a 0.04 increase in PW boundary f1 score, compared with purely BERT pretraining. This shows that contrastive pretraining is more effective in unseen speaker scenario due to its generalization capability learned from large pretrained data.

2) SSWP contributes to a 0.06-0.11 increase in PW f1 score and 0.08-0.11 PPH f1 score under all three scenarios. SSWP is the only component that achieves significant gains on PPH boundary classification. This is because SSWP, as a linguistically motivated design, is a more appropriate text-speech pair for prosody representation and prosody annotation tasks.

3) The bi-LSTM classification network is useful, especially in low resource scenario, contributing to a 0.14 increase in PW boundary f1 score, and a 0.05 in PPH boundary f1 score. The contextual information included in the recurrent network makes the overall system more robust to smaller amounts of data.

4.3. Subjective Evaluation

We also conduct an AB Preference evaluation on an open-source corpus of an unseen speaker, LJSpeech [37], using the `Multi-modal Baseline` and `Proposed Work` to annotate the prosodic boundaries of the entire corpus automatically. Both systems are trained using Conformer-FastSpeech2 [31] and Hifi-GAN [38], with a 16-dimension embedding encoding the prosodic boundary labels concatenated to the phone embedding as input. All models are trained for 200k steps with a batch size of 32. 30 native speakers are asked to compare the 30 utterances from each system and choose the one they prefer. The results are shown in Table 5. Our proposed work achieves a **7.58%** improvement in AB Preference compared with the previous benchmark. This demonstrates slight improvement in naturalness of the synthesised audio by our approach, in addition to the higher annotation accuracy for developing controllable TTS systems demonstrated in previous sections. Demos are openly available at `https://jzmzhong.github.io/Automatic-Prosody-Annotator-With-SSWP-CLAP/`

Table 5: *AB Preference results with 95% confidence interval calculated for TTS systems trained on a corpus annotated by different prosody annotators. (p-value = 0.033)*

Prosody Annotation System	AB Preference (%)
Multi-modal Baseline	46.21 ± 3.33
Proposed Work	53.79 ± 3.33

5. Conclusions & Future Work

In this paper, we propose a novel two-stage pipeline that achieves SOTA performance on automatic annotation of English prosodic boundaries. With the design of contrastive text-speech pretraining of SSWP pairs, text and audio encoders learn richer prosody information. We also use aligned feature fusion and a sequence classification network to improve prosody annotation with contextual information. In the future, we will investigate the proposed work in cross-lingual scenarios and on other prosodic features. We will also investigate the possibility of developing a unified TTS annotation tool covering phonetic information (out-of-vocabulary words, heteronyms, etc.) and other speech information (emotion, style, accent, etc.).

6. References

[1] Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li, X. Tan, J. Li, L. He, and S. Zhao, "DelightfulTTS: The Microsoft Speech Synthesis System for Blizzard Challenge 2021," *arXiv preprint arXiv:2110.12612*, 2021.

[2] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] Y. Zhou, C. Song, J. Li, Z. Wu, Y. Bian, D. Su, and H. Meng, "Enhancing Word-Level Semantic Representation via Dependency Structure for Expressive Text-to-Speech Synthesis," in *Proc. Interspeech*, 2022, pp. 5518–5522.

[4] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, "PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS," in *Proc. Interspeech*, 2021, pp. 151–155.

[5] G. Zhang, K. Song, X. Tan, D. Tan, Y. Yan, Y. Liu, G. Wang, W. Zhou, T. Qin, T. Lee, and S. Zhao, "Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech," in *Proc. Interspeech*, 2022, pp. 456–460.

[6] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, "Phoneme-Level Bert for Enhanced Prosody of Text-To-Speech with Grapheme Predictions," in *Proc. ICASSP*, 2023, pp. 1–5.

[7] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *Proc. ICLR*, 2020.

[8] Y. Ren, M. Lei, Z. Huang, S. Zhang, Q. Chen, Z. Yan, and Z. Zhao, "Prosospeech: Enhancing prosody with quantized vector pre-training in text-to-speech," in *Proc. ICASSP*, 2022, pp. 7577–7581.

[9] Z. Ye, R. Huang, Y. Ren, Z. Jiang, J. Liu, J. He, X. Yin, and Z. Zhao, "CLAPSpeech: Learning Prosody from Text Context with Contrastive Language-Audio Pre-Training," in *Proc. ACL*, 2023, pp. 9317–9331.

[10] M. Babiański, K. Pokora, R. Shah, R. Sienkiewicz, D. Korzekwa, and V. Klimkov, "On Granularity of Prosodic Representations in Expressive Text-to-Speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 892–899.

[11] Y. Ren, J. Liu, and Z. Zhao, "PortaSpeech: Portable and High-Quality Generative Text-to-Speech," *Proc. NeurIPS*, vol. 34, pp. 13 963–13 974, 2021.

[12] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[13] Z. Dai, J. Yu, Y. Wang, N. Chen, Y. Bian, G. Li, D. Cai, and D. Yu, "Automatic Prosody Annotation with Pre-Trained Text-Speech Model," in *Proc. Interspeech*, 2022, pp. 5513–5517.

[14] X. Yuan, R. Feng, and M. Ye, "Low-Resource Mongolian Speech Synthesis Based on Automatic Prosody Annotation," *arXiv preprint arXiv:2211.09365*, 2022.

[15] A. Rosenberg, "AutoBI - A Tool for Automatic ToBI Annotation," in *Proc. Interspeech*, 2010, pp. 146–149.

[16] Y. Zou, S. Liu, X. Yin, H. Lin, C. Wang, H. Zhang, and Z. Ma, "Fine-Grained Prosody Modeling in Neural Speech Synthesis Using ToBI Representation," in *Proc. Interspeech*, 2021, pp. 3146–3150.

[17] B. Stephenson, L. Besacier, L. Girin, and T. Hueber, "BERT, Can HE Predict Contrastive Focus? Predicting and Controlling Prominence in Neural TTS Using a Language Model," in *Proc. Interspeech*, 2022, pp. 3383–3387.

[18] N. Sanders and K. Richmond, "Invert-Classify: Recovering Discrete Prosody Inputs for Text-to-Speech," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[19] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical Representation and Estimation of Prosody Using Continuous Wavelet Transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[20] R. Sloan, S. S. Akhtar, B. Li, R. Shrivastava, A. Gravano, and J. Hirschberg, "Prosody Prediction from Syntactic, Lexical, and Word Embedding Features," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 269–274.

[21] X. Chen, C. Song, Y. Zhou, Z. Wu, C. Chen, Z. Wu, and H. Meng, "A Character-Level Span-Based Model for Mandarin Prosodic Structure Prediction," in *Proc. ICASSP*, 2022, pp. 7602–7606.

[22] S. Ananthakrishnan and S. S. Narayanan, "Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 216–228, 2008.

[23] V. K. Rangarajan Sridhar, S. Bangalore, and S. S. Narayanan, "Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 797–811, 2008.

[24] Y. Qian, Z. Wu, X. Ma, and F. Soong, "Automatic Prosody Prediction and Detection with Conditional Random Field (CRF) models," in *2010 7th International Symposium on Chinese Spoken Language Processing*, 2010, pp. 135–138.

[25] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[27] Z. Li, B. Tang, X. Yin, Y. Wan, L. Xu, C. Shen, and Z. Ma, "PPG-Based Singing Voice Conversion with Adversarial Representation Learning," in *Proc. ICASSP*, 2021, pp. 7073–7077.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[29] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in *Proc. ICASSP*, 2023, pp. 1–5.

[30] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP Learning Audio Concepts from Natural Language Supervision," in *Proc. ICASSP*, 2023, pp. 1–5.

[31] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent Developments on Espnet Toolkit Boosted By Conformer," in *Proc. ICASSP*, 2021, pp. 5874–5878.

[32] E. O. Selkirk, "On Prosodic Structure and its Relation to Syntactic Structure," *Nordic Prosody II*, 1980.

[33] K. Steinhauer, "Electrophysiological Correlates of Prosody and Punctuation," *Brain and language*, vol. 86, no. 1, pp. 142–164, 2003.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-scale Weak Supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[36] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[37] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[38] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *Proc. NeurIPS*, vol. 33, pp. 17 022–17 033, 2020.