Subphoneme-Level PRESENT: PRosody Editing without Style Embeddings or New Training

Perry Lam, Huayun Zhang, Nancy F. Chen, Berrak Sisman, Dorien Herremans

***Abstract*—Current strategies for achieving fine-grained prosody control in speech synthesis entail the extraction of additional style embeddings or adoption of more complex architectures. To enable zero-shot application of pretrained text-to-speech (TTS) models, we present PRESENT, which exploits explicit prosody prediction in FastSpeech2-based models by modifying the inference process directly. This approach empowers us to attain subphoneme-level control, a first in this field, mitigating artefacts stemming from extended phoneme durations. Although prosodic modifications achieved solely through inference parameters may not reach the same quality levels as those attained through digital signal processing techniques, they can improve the prosody of questions. To evaluate the effectiveness of our subphoneme-level control, we conducted experiments using a JETS model exclusively trained on English LJSpeech data to generate Mandarin, a tonal language. We attain 34.4% Character Error Rate (CER) at hanzi level and 17.2% CER at pinyin-level, a state-of-the-art performance.**

All our code[1] and audio samples are available online[2].

***Index Terms*—speech synthesis, prosody, computational paralinguistics, zero-shot, language transfer**

I. Introduction

RECENT neural text-to-speech (TTS) models have approached human-like naturalness in read speech. However, attaining similar expressiveness levels remains a challenge. A growing body of research aims to add and control speech prosody variations, progressing from digital signal processing (DSP) methods to style and emotion embeddings built into TTS architectures or even entire models to extract and transfer prosody.

On the waveform level, prosody control can be achieved through operations like time-stretching and pitch-shifting. DSP methods such as TD-PSOLA [1] and WORLD [2], despite their known artifacts, are still widely applied due to their speed and ease of use. Remarkably, they can perform as effectively as neural approaches like Controllable LPCNet [3].

In contrast, expressive TTS systems [4] allow the user to specify a style or emotion label during inference. Traditional methods rely on handcrafted rules specific to each emotion to directly manipulate the speech signal or statistical parameters during voice generation [5]. More recently, TTS models incorporate style or emotion information by extracting a reference embedding that represents the prosody or emotion from labelled audio, and adding it to the model encoder. This can be combined with a style bank for smooth style variation, such as in Global Style Tokens [6]. Further extensions include phoneme-level prosody control and hierarchical autoencoders to ensure coherence over the whole utterance [7].

When explicit labels are not available but a source audio sample is available, prosody transfer models such as [8] can employ a reference encoder to create prosodic representations for style transfer onto the input text. However, one limitation is the potential entanglement of speaker and text information with the prosody embedding, as highlighted in [9], and current research, such as [10], are dedicated to addressing and mitigating this issue.

All of these approaches, however, require extra model components and/or further training. Therefore, to combine the simplicity of DSP methods with the naturalness of neural speech generation, we empower users to directly control prosody using the input text and inference parameters without the need for any fine-tuning or architectural modifications. We contribute significantly in the following three areas:

- Extraction of prosodic effects from text, such as extended duration in “A looooong time" or the intonation variations in questions like “What was that?";
- Attainment of subphoneme-level control, achieved by subdividing phonemes and applying custom pitch and energy over the subdivisions, which mitigates artefacts related to unusually long phonemes;
- Zero-shot language transfer with no target-language audio at all, by relying solely on linguistic knowledge and modifying the inference method of any TTS model with explicit duration, pitch, and energy (DPE) predictions.

Though our primary goal is to explore the limits of editing inference-time prosody predictions, in doing so, we achieve state-of-the-art results in the challenging task of zero-shot language transfer.

The rest of this paper is organized as follows: Section 2 summarizes relevant research, Section 3 describes our approach, Section 4 lists our experiment results and Section 5 concludes our paper.

II. Related Work

Based on our main contributions, we divide the related work into the broad categories of (1) speech effect tagging, (2) fine-grained prosody control, and (3) zero-shot language transfer.

Submitted on 6 Nov, 2023.

Perry Lam and Dorien Herremans are with Singapore University of Technology & Design, Singapore 487372 (e-mail: perry_lam@mymail.sutd.edu.sg and dorien_herremans@sutd.edu.sg).

Huayun Zhang and Nancy F. Chen are with the Institute of Infocomm Research, A*STAR, Singapore 138632 (e-mail: Zhang_Huayun@i2r.a-star.edu.sg and nfychen@i2r.a-star.edu.sg).

Berrak Sisman is with the University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: berrak.sisman@utdallas.edu).

[1]https://github.com/iamanigeeit/present

[2]https://present2023.web.app/

A. Speech Effects Tagging

Text-based methods for manipulating speech can be categorized into explicit and implicit forms. Explicit speech descriptors have been integrated into the industry standard Speech Synthesis Markup Language (SSML) over the past two decades [11], encompassing speaker and prosody tags, including parameters like gender, emphasis, pitch, duration, and volume. The precise implementation of these tags is left to the Text-to-Speech (TTS) synthesis processor. While these tags enjoy widespread support in commercial TTS systems, there has been relatively limited published research on SSML, even though there have been notable introductions of TTS models with new style tags, as demonstrated in [12].

Implicit methods, on the other hand, establish connections between prosodic features and text, such that a sentence like "this is interesting!" would sound excited. Typically, this means that the text embeddings from a language model (most often BERT) are used as input either at the subword [13] [14] or phoneme level [15] [16]. However, due to their inherent limitations in customizing prosody changes, more recent work, as inspired by advancements in computer vision and language processing, now allows users to input a natural-language style prompt like “sighing with helpless feeling" to generate prosodic output, as exemplified in [17].

B. Fine-grained Prosody Control

As utterance-level styles are now commonplace, research has shifted to controlling prosody at the phoneme level. Since acceptable prosodies are obtained by learning and sampling from a variational latent space, hierarchical variational auto-encoders (VAEs) [18] can achieve fine prosodic gradations, down to the syllable, phone or even frame level [7].

Alternatively, others use phone-level DPE for interpretable prosody control. This was the approach of earlier research [19], but to improve output naturalness, [20] and [21] used k-means clustering on duration and pitch for each speaker, and kept the resulting centroids as discrete prosody tokens. This allows the tokens to be substituted at inference time to customize prosody, while decoding with a prosody attention module ensures information flows to the output. Meanwhile, since the advent of explicit DPE models like FastSpeech2 [22], models like [23] and [24] have extra modules attached that accept emotional dimensions (valence, arousal, dominance) that feed into phone-level DPE predictors, allowing for continuous emotion control.

C. Zero-Shot Language Transfer

While multilingual TTS models have existed for some time, they rely on large multilingual corpora, which disadvantages lower-resourced languages. Transfer learning [25] [26] and data balancing [27] techniques have been employed, but these still require at least some audio data. With only International Phonetic Alphabet (IPA) transcriptions in the target language, [28] proposed using IPA phonological features to extend existing models on unseen phonemes, whereas two very recent large models have proposed zero-shot TTS with only text data available in the target language.

The first model, VALL-E X [29], uses AudioLM codec codes as acoustic tokens in place of mel spectrograms as intermediate features, and treats the cross-lingual TTS model as a massive language model (LM) that can be trained with self-supervised masking. Given a speech sample in the source language, plus source and target language phoneme sequences, it extracts the source acoustic tokens from the speech sample and the LM predicts the target acoustic tokens. Since the acoustic tokens contain speaker, recording conditions, and subphoneme information, the decoder can reconstruct the waveform for the target language in the source speaker’s voice.

The second model, ZM-Text-TTS [30], also uses masked multilingual training, but on IPA / byte tokens and raw text. The pretraining results in a language-aware embedding layer that is fed to a conventional multilingual TTS system for training with seen languages, and the model can accept IPA / byte tokens for unseen languages during inference. Nevertheless, VALL-E X is not publicly available, and ZM-Text-TTS does not account for prosody in language transfer.

III. PROPOSED METHOD

PRESENT offers a versatile approach to extract inference parameters and merge them with explicit duration, pitch, and energy predictions to generate variations in pronunciation and prosody, all without requiring additional modules or fine-tuning. The specific method of parameter extraction and integration is adaptable to the task at hand and can be customized by the user. An overview of the entire process is in Fig 1.

A. Text Alignment

We preprocess the input text to capture common dialogue features such as CAPS or *asterisks* for emphasis, repeaaaaated letters or ti~~lde for long phonemes, and special characters like underscores and carets or questions for tone modification. As TTS systems usually rely on phoneme input, we align the text to phonemes so that the duration, pitch and energy changes can be applied at the correct positions. Despite the availability of grapheme-to-phoneme systems (G2P), G2P alignment systems are outdated and notably lacking in Python implementations. As a result, we have developed our own aligner, complete with a constrained ruleset and search for the least-cost path if a constraint-satisfying alignment cannot be found. This is similar to the “Phonetic Alignment" and “IP Alignment" methods in [31] and more details can be found in our code.

B. Subphoneme-level Control

Pitch-predicting TTS models produce a single pitch value for each phoneme. However, to achieve tone contour effects such as the rising-falling pitch in “Suuuuure!", the phoneme should be divided and separate pitches assigned to each subphoneme. Thus, we repeat the encoder output $\boldsymbol{h}$ for the divided phoneme (the first yellow box of `encoder_hs` in Fig 1) , which differs from simply repeating the phonemes for inference (as that would generate different $\boldsymbol{h}$ and possibly cause it to be pronounced multiple times).

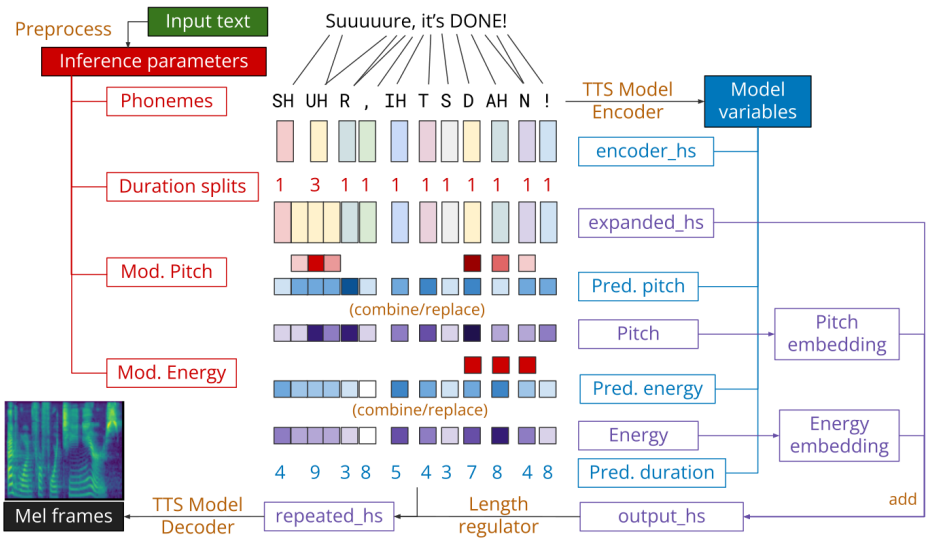

Fig. 1. Overview of inference modification in PRESENT. Red items represent PRESENT outputs, blue items are the outputs from the FS2-based model, purple are the combined outputs, and gold are operations. White-filled text boxes are tensors.

One evident use case for pitch effects pertains to questions. While humans can clearly perceive a question via prosody, TTS systems still lack proper intonation. For English question prosody, which is known for its complexity, a solution was introduced in [32], enabling users to choose from a range of pretrained prosody templates. However, to maintain simplicity and avoid adding the complexity of language models, we follow the prosodic analysis of [33], applying a low-to-high accent on the locus of interrogation and the final word of the question to convey question intonation.

C. Zero-Shot Language Transfer

Another critical test of our subphoneme tone contour approach is whether it can model tonal languages. Therefore, we get an English JETS [34] TTS model trained on only LJSpeech to generate Mandarin, without any Mandarin data. This is extremely challenging as LJSpeech is relatively monotonous single-speaker read speech, and American English differs radically from Mandarin phonologically. We approach this language transfer problem by (1) creating a Pinyin-to-ARPA conversion table and (2) splitting each vowel into subphonemes and interpolating tone contours across each syllable.

Our Pinyin-to-ARPA converter corrects phoneme mismatches by assigning pinyin with no corresponding ARPA realization to ARPA combinations with varying or zero duration, which reflects the power of editable durations. For example, pinyin ⟨d⟩ is given ARPA ⟨T D⟩ with 0 duration on ⟨D⟩ to prevent voicing onset, while pinyin ⟨t⟩ results in ARPA ⟨T HH⟩ to ensure aspiration. Pinyin ⟨a⟩, which is more open than ⟨AH⟩ but more closed than ARPA ⟨AA⟩, resolves to ⟨AH AA⟩ with 0 duration on ⟨AA⟩. Similarly, we represent pinyin ⟨ü⟩ with ARPA ⟨UW IY⟩ with 0-duration ⟨UW⟩ coarticulating with ⟨IY⟩ to approximate the rounded Mandarin vowel. The full conversion matrix is available in our source code and most resembles MPS-II romanization.

We then map Mandarin tones on the five-point tonal scale to normalized pitch values between $[-2.0, +2.0]$ according to Table I.

TABLE I
TONE-PITCH MAPPING.

Tone	1	2	3	4	5
Contour	55	25	212	52	-
Pitch	+2, +2	-1, +2	-1, -2, -1	+2, -1	0

The tone contour is subsequently applied across the given syllable, treating both initial and coda as single phonemes. For example, Table II demonstrates how pinyin ⟨tián⟩ maps to ARPA ⟨T HH Y EH N⟩. We then smooth pitch transitions across syllables to avoid any abrupt pitch changes.

TABLE II
EXAMPLE OF ARPA-PINYIN MAPPING.

ARPA	T	HH	Y	EH			N
Duration	×1	×0.5	×1	×1			×1
Subphonemes	T	HH	Y	EH	EH	EH	N
Pitch	-1	-1	-0.4	+0.2	+0.8	+1.4	+2

As Mandarin is a syllable-timed language, we keep the duration of the ×1 syllable nuclei constant, with the neutral tone at half duration. Additionally, we leverage `pywordseg` to segment Mandarin text and introduce brief pauses between words, to improve enunciation.

IV. EXPERIMENTS

We conducted our experiments using the ESPnet [35] toolkit for reproducibility. Our baseline was the publicly released JETS model pretrained on LJSpeech, known for achieving state-of-the-art naturalness. All experiments were conducted with PsyToolkit [36] [37] and 15 responses were received.

A. Prosodic Effects

We assess the quality of our prosodic effects with Mean Opinion Score (MOS) on (1) sentences with extra-long phonemes and (2) questions. For (1), we created our own sentences with words to be lengthened in the absence of a

dataset. To standardize comparisons, we generate the sentence with JETS, then apply 4× time-stretching in different ways: by TD-PSOLA (as implemented in Praat [38] Parselmouth), by forcing JETS to generate with increased duration, and by using PRESENT (JETS with increased duration and pitch contour). For (2), we took the first 10 dialogues from the DailyTalk dataset [39] and extracted the first single-sentence question from each of them, making 10 questions in total. We report the MOS for ground truth audio from DailyTalk, unaccented JETS-generated audio, and PRESENT-accented audio.

TABLE III
PROSODIC EFFECT MOS.

	Extra-long	**Questions**
Ground Truth (DailyTalk)	-	4.46
TD-PSOLA	**3.15**	-
JETS	3.08	3.73
PRESENT	2.85	**3.92**

Our results in Table III reinforce the continued relevance of TD-PSOLA as noted in [40]. While PRESENT could reduce some artefacts in extended phonemes, we surmise that the added pitch contours made the speech sound unnatural, and we ask readers to listen to the audio samples. Nevertheless, in line with our expectations, applying subphoneme pitch contours to questions improved in their overall naturalness.

B. Zero-Shot Language Transfer

We evaluated the ability of the English-only JETS model to synthesize intelligible Mandarin speech by synthesizing speech based on the AISHELL-3 test set transcripts. As a baseline, we use the pretrained IPA multilingual model (trained on 7 European languages) from ZM-Text-TTS [30] to generate Mandarin. To do this, we convert pinyin transcripts to IPA and use the best-approximation phoneme when a Mandarin phoneme does not exist in the pretrained IPA symbol set, and run inference without any language or speaker embeddings. We then input the synthesized audio into the state-of-the-art pretrained FunASR Paraformer automatic speech recognition framework [41], using the Chinese-only `aishell2-vocab5212` model to avoid English words in the transcript. For comparison, we list the character error rate (CER) for Paraformer transcriptions of ground-truth audio, PRESENT-generated samples, samples generated without subphoneme tone contours applied, and the baseline (Table IV).

TABLE IV
ENGLISH-TO-MANDARIN LANGUAGE TRANSFER RESULTS. CER HERE IS AT HANZI LEVEL, I.E. EACH CHINESE IDEOGRAM IS ONE CHARACTER. "PERFECT" IS THE PROPORTION OF PERFECT TRANSCRIPTIONS. NOTE THAT THE BASELINE HAS OVER 100% CER DUE TO MANY INSTANCES OF GENERATING MORE SYLLABLES THAN GROUND TRUTH.

	% **Hanzi CER**	% **Perfect**
Ground Truth	2.1	83.6
PRESENT	**34.4**	**14.5**
PRESENT (no tone contours)	75.9	2.0
Baseline (ZM-Text-TTS)	105.5	0.0

The dramatic improvement in CER of PRESENT compared to both ZM-Text-TTS and PRESENT without tone contours applied demonstrates the effectiveness of our subphoneme-level prosody control.

ZM-Text-TTS [30] uses the 8 European languages of the CSS10 multilingual dataset [42]. However, their CER numbers cannot be compared directly to Table IV, because phoneme/grapheme-based CER would be lower than the syllable-based CER for Mandarin (e.g. a `la` to `lo` change is 50% CER at grapheme level but 100% at syllable/hanzi level). Moreover, due to homophony in Mandarin, the same syllable may map to a different character, inflating hanzi CER. Thus, for fair comparison against Latin-based orthography, we romanize Mandarin transcripts with `pypinyin` and report pinyin-level CER in Table V, with tone counting as one character (i.e. `ping1` versus `ping2` would be 20% CER). The CER ranges shown for ZM-Text-TTS consist of both fully zero-shot TTS (higher CER) and text-seen zero-shot TTS (where raw text is available in the target language for multilingual pretraining, which results in lower CER).

TABLE V
CER COMPARISON AFTER CONVERTING MANDARIN TRANSCRIPTS TO PINYIN. NUMBERS FOR GERMAN, HUNGARIAN AND SPANISH ARE FROM THE UNSEEN-LANGUAGE SETTING.

	ZM-Text-TTS				**PRESENT**
Source	6 European langs		7 European langs		English
Target	German	Hungarian	Spanish	Mandarin pinyin	
CER	28.0–38.8	50.1–52.6	13.3–44.4	75.0	**17.2**

Our CER is state-of-the-art, even when compared to European-language transfer scenarios. This is a notable achievement, especially considering the two key challenges we faced: (1) the vast typological differences between English and Mandarin, when contrasted with the similarities among European languages, and (2) the fact that we merely use a single off-the-shelf English-only model to perform zero-shot language transfer with no further training.

V. CONCLUSIONS

We have introduced PRESENT, a novel approach that explores the limits of using only duration, pitch and energy predictions in a single-speaker English-only JETS model, without the need for additional embeddings or training. Our research explored various tasks, including creating prosodic effects and synthesizing speech in an unseen tonal language. Among these tasks, we found that the subphoneme control mechanism we introduced is particularly well-suited for improving question prosody and enabling zero-shot language transfer to tonal languages. Our language transfer method from English to Mandarin achieves state-of-the-art CER results compared to European language transfer at pinyin level. Furthermore, the phoneme conversion and tone contour techniques we develop could pave the way for simple accented speech generation (as we produce English-accented Mandarin), or TTS solutions for hundreds of minority languages within the Mainland Southeast Asian linguistic area. Many of these languages are only recorded in phonetic transcriptions, presenting an exciting opportunity for linguistic diversity and accessibility in TTS.

REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[2] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[3] M. Morrison, Z. Jin, N. J. Bryan, J.-P. Caceres, and B. Pardo, "Neural pitch-shifting and time-stretching with controllable lpcnet," *arXiv preprint arXiv:2110.02360*, 2021.

[4] J. Pitrelli, R. Bakis, E. Eide, R. Fernandez, W. Hamza, and M. 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.

[5] F. Burkhardt and N. Campbell, "Emotional speech synthesis," in *The oxford handbook of affective computing*, R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Eds. Oxford: Oxford University Press, 2014, ch. 20, pp. 286–294.

[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] 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 *International Conference on Machine Learning*. PMLR, 2019, pp. 3331–3340.

[8] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 970–10 983, 2022.

[9] A. T. Sigurgeirsson and S. King, "Do prosody transfer models transfer prosody?" in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] J. Zaïdi, H. Seuté, B. van Niekerk, and M.-A. Carbonneau, "Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 4591–4595.

[11] Z. W. Shuang and D. Burnett, "Speech synthesis markup language (SSML) version 1.1," W3C, W3C Recommendation, Sep. 2010, https://www.w3.org/TR/2010/REC-speech-synthesis11-20100907/.

[12] M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim, and N. S. Kim, "Expressive Text-to-Speech Using Style Tag," in *Proc. Interspeech 2021*, 2021, pp. 4663–4667.

[13] S. Ammar Abbas, T. Merritt, A. Moinet, S. Karlapati, E. Muszynska, S. Slangen, E. Gatti, and T. Drugman, "Expressive, Variable, and Controllable Duration Modelling in TTS," in *Proc. Interspeech 2022*, 2022, pp. 4546–4550.

[14] L. Chen, Y. Deng, X. Wang, F. K. Soong, and L. He, "Speech bert embedding for improving prosody in neural tts," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6563–6567.

[15] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, "Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] 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*, 2021, pp. 151–155.

[17] D. Yang, S. Liu, R. Huang, G. Lei, C. Weng, H. Meng, and D. Yu, "Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt," *arXiv preprint arXiv:2301.13662*, 2023.

[18] 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.

[19] 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.

[20] P. Tsiakoulis, "Improved prosodic clustering for multispeaker and speaker-independent phoneme-level prosody control," in *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings*, vol. 12997. Springer Nature, 2021, p. 112.

[21] N. Ellinas, M. Christidou, A. Vioni, J. S. Sung, A. Chalamandaris, P. Tsiakoulis, and P. Mastorocostas, "Controllable speech synthesis by learning discrete phoneme-level prosodic representations," *Speech Communication*, vol. 146, pp. 22–31, 2023.

[22] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR*, 2021.

[23] S. Sivaprasad, S. Kosgi, and V. Gandhi, "Emotional Prosody Control for Speech Generation," in *Proc. Interspeech 2021*, 2021, pp. 4653–4657.

[24] S. Kosgi, S. Sivaprasad, N. Pedanekar, A. Nelakanti, and V. Gandhi, "Empathic machines: using intermediate features as levers to emulate emotions in text-to-speech systems," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 336–347.

[25] T. Nekvinda and O. Dušek, "One Model, Many Languages: Meta-Learning for Multilingual Text-to-Speech," in *Proc. Interspeech 2020*, 2020, pp. 2972–2976.

[26] K. Azizah, M. Adriani, and W. Jatmiko, "Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on low-resource languages," *IEEE Access*, vol. 8, pp. 179 798–179 812, 2020.

[27] J. Yang and L. He, "Towards Universal Text-to-Speech," in *Proc. Interspeech 2020*, 2020, pp. 3171–3175.

[28] M. Staib, T. H. Teh, A. Torresquintero, D. S. R. Mohan, L. Foglianti, R. Lenain, and J. Gao, "Phonological Features for 0-Shot Multilingual Speech Synthesis," in *Proc. Interspeech 2020*, 2020, pp. 2942–2946.

[29] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

[30] T. Saeki, S. Maiti, X. Li, S. Watanabe, S. Takamichi, and H. Saruwatari, "Learning to speak from text: Zero-shot multilingual text-to-speech with unsupervised text pretraining," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. ijcai.org, 2023, pp. 5179–5187. [Online]. Available: https://doi.org/10.24963/ijcai.2023/575

[31] S. Jiampojamarn and G. Kondrak, "Phoneme alignment: An exploration," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 780–788.

[32] J. Lee, J. Y. Lee, H. Choi, S. Mun, S. Park, J.-S. Bae, and C. Kim, "Into-tts: Intonation template based prosody control system," *arXiv preprint arXiv:2204.01271*, 2022.

[33] N. Hedberg and J. M. Sosa, "The prosody of questions in natural discourse," in *Speech Prosody 2002, International Conference*, 2002.

[34] D. Lim, S. Jung, and E. Kim, "JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech," in *Proc. Interspeech 2022*, 2022, pp. 21–25.

[35] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7654–7658.

[36] G. Stoet, "Psytoolkit: A software package for programming psychological experiments using linux," *Behavior Research Methods*, vol. 42, pp. 1096–1104, 2010.

[37] ——, "Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.

[38] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2011.

[39] K. Lee, K. Park, and D. Kim, "Dailytalk: Spoken dialogue dataset for conversational text-to-speech," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[40] M. Morrison, L. Rencker, Z. Jin, N. J. Bryan, J.-P. Caceres, and B. Pardo, "Context-aware prosody correction for text-based speech editing," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7038–7042.

[41] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition," in *Proc. Interspeech 2022*, 2022, pp. 2063–2067.

[42] K. Park and T. Mulc, "CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages," in *Proc. Interspeech 2019*, 2019, pp. 1566–1570.