
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1557–1563, DOI 10.12694/scpe.v25i3.2710

DEEP MACHINE LEARNING-BASED ANALYSIS FOR INTELLIGENT PHONETIC LANGUAGE RECOGNITION

YUMEI LIU* AND QIANG LUO†

Abstract. Modern speech generating systems can produce results that are almost as visually realistic as actual sounds. They still require further production management. This research presents a paradigm for managing prosodic output using explicit, unambiguous, and understandable parameters. We utilize this strategy to emphasize key words and provide a variety of architectural possibilities based on a richness of labelled resources. In an objective voice, we compare the options for producing data with or without labels. We assess them using listening tests that demonstrate our ability to retain the same level of naturalness while effectively attaining regulated concentration over a specific area.

Key words: Prosody management, machine learning, speech analysis, lexical focus.

1. Introduction. In today's digital age, the rapid development of human-computer interaction and natural language processing technology has led to widespread attention to Automatic Speech Recognition (ASR). Intelligent speech recognition systems have potential applications, ranging from virtual assistants and smart home controls to healthcare and education, providing users with convenient and efficient voice interaction methods. The progress in this field not only provides better experiences for individual users, but also provides more innovation and business opportunities for enterprises and organizations. The most advanced architectures now available offer high-quality results that typically approach or equal what is expected of natural language [7]. Apart from the high quality, these models contain a number of enticing features. They may jointly define multiple waveform properties, allowing correlations between them to be observed . Furthermore, they abandon typical pipeline designs in favour of a loosely connected, unified approach, which is desirable when some pipeline modules (for example, text processing in a foreign language) are difficult to construct. However, they do have some well-known drawbacks, including interpretability issues (it can be difficult to determine which parts of the model are responsible for what functions), controllability issues (it can be challenging to intervene in the model to influence some aspects of the synthesis, which is frequently desired, such as when providing SSML support), and potential instability issues (minor deviations at inference time can become worsened and gen. By adding methods to the architecture that the user can use to modify a particular output property, this work tackles the controllability issue [20]. Although usability factors are not the main emphasis of this project, we support a set of characteristics that will enable these options to be available to the system's end user.

- **Interpreting**: The listener should be able to hear and recognize the impact of a control change (for instance, whether speech is sluggish quicker, better-pitched, or seems more joyful, etc.) [11].
- **Monotonicity**: An aesthetic with sensory impacts that alter inversely when the user changes the knob feels better natural and is simpler to tune [2].
- **Low-dimensionality**: The user shouldn't need to alter many parameters to change the outcome. The model should be able to offer a low-dimensional customizable representation or step in and complete up defaults to take care of the user's work [5].

Disentanglement: While this may be challenging due to the multiple ways in which distinct expression variables interact, controlling the output along relatively separate (perceptual) dimensions is made simpler by a set of controls that are more dissociated from one another [17]. Speed and quantity, for example, may be tweaked independently without having to go back and alter a previously adjusted variable. We analyse the implementation and controllability of constrained lexical emphasis as a case study for those stated previously .

*Chongqing City Vocational College, Chongqing, 402160, China (Corresponding Author, YumeiLiu7@126.com)
†Chongqing Creation Vocational College, Chongqing, 402160, China (QiangLuo85@163.com)

We want to produce an expressive amount of importance that is distinct from prosodic accentuation by using a "neutral" wide focus. Consider how the phrase that has become a catchphrase responds to the circumstances below. In these instances, the speaker emphasises the focus area by separating the target object from its surroundings more clearly [12]. To produce flexible and suitable synthetic speech, control over the output expression must exist independently of spoken text. Since significant non-textual speech variance is rarely marked, output control must be learned unsupervised. In this research, we thoroughly investigate techniques for statistical speech synthesis unsupervised learning of control. For instance, we demonstrate how some auto encoder models can interpret standard unsupervised training techniques as variational inference. These new probabilistic interpretations' ramifications are examined. It encourages the potential of unsupervised learning to provide output control in speech synthesis in general [8].

Amortized inference-based methods are promising for upcoming applications since they provide comparable performance to existing heuristics, making training and latent-variable inference easier. We can force some latent variables to adopt consistent and understandable purposes by providing partial supervision to some of them, which was previously impossible with completely unsupervised TTS models. With as little as 1% (30 minutes) of care, our model can accurately find and regulate crucial but rarely tagged speech characteristics, such as effect and speaking rate [15].

2. Literature Survey. We can consistently and reliably learn to regulate specified parts of prosody, in contrast to earlier wholly unsupervised techniques. Any latent aspect of speech, continuous or discrete, for which a moderate amount of labelling can be collected, can be used with our method. When precise duration labels are unavailable or sparse in the training data, the proposed model can be trained explicitly with duration labels or unsupervised or semi-supervised using a fine-grained variational auto-encoder . When trained and wholly supervised, the proposed model slightly beats Tacotron 2 on naturalness [13]. The suggested model outperforms Tacotron 2 on naturalness with unsupervised or semi-supervised duration modelling while still much more resilient on over-generation and equivalent on under-generation.

At the time of inference, the duration predictor additionally offers per-phoneme and utterance-wide duration control. This study introduced the Non-Attentive Tacotron, which considerably exceeded Tacotron 2 in terms of robustness as measured by the unaligned duration ratio and word deletion rate while outperforming Tacotron 2 in terms of naturalness. Tacotron 2's attention algorithm was replaced with Gaussian upsampling and an explicit duration predictor to achieve this. We also demonstrated that the duration predictor could be used to change both the utterance's overall pacing and the rate at which individual syllables are spoken [18].

The technique works with both expanded state sequences—each corresponding to a single feature frame—and state sequences with defined durations. We also give a thorough examination blended sample' phonetic composition. The evaluation incorporates phonetically motivated, gradual, and universally applicable phonological processes and input-switch rules, encompassing the dialects' historically divergent phonological evolution against the standard language. We describe an expanded technique that uses a step function for input-switch practices while linearly interpolating phonological processes [10].

Our investigation shows that phonological knowledge of this kind improves dialect speakers' capacity to judge the dialect authenticity of synthesised speech. Our methods can be utilised to alter voice output systems because progressive alterations between kinds are a common occurrence. For state-level interpolation, it locates HSMM-state mappings using DTW. One feature frame is produced for each state using either of two techniques for dealing with state durations: either continuing with the unexpanded states or increasing every instance with a length of N to N states with an interval of 1 [4].

Our findings imply that DTW's linear interpolation of its mapped HSMM states was fair. The machine translation component of a comprehensive interpolation system would also translate a standard variety into dialect. We would get input switch rules for words from this component, which may also produce syntactic modifications. Phonetic criteria must be used to derive rules for phonemes [16].

3. Materials and Methods.

3.1. Design. The model has been enhanced with decoder-to-facilitator components. 25 January 2021 saw the addition of controls and increased stability during decoding. This series-to-sequence model generates an auditory spectroscopy (eds-prosodic model that is then fed to an asynchronously instructed, LPC-Net-based

brain vocoder in order to generate high-quality samples in real-time.

- Emphasis anchoring (A) is a basic verification method based on binary indicative features, which is used to derive the emphasis focus in discourse. The basic principle of this method is to determine the position of emphasis focus by analyzing specific features in discourse, namely binary indicative features. This process can be regarded as an equation with the aim of verifying the existence and position of the focal point. In short, emphasis anchoring is a method of analyzing discourse structure by identifying specific binary indicative features to determine the focus of emphasis in discourse. This method helps to gain a more detailed understanding of the role and expression of emphasis in language [19].
- A front-end programming generator (C) that utilizes bidirectional short-term memory (Bi-LSTM) layers and convolutional layers to encode the combined embeddings from (A) and (B).
- To aid with training in a setting with several speakers, an overall phrases-level speaker embedding (D) is distributed throughout the episode.

The Decoder is an autoregressive network that modifies the fundamental architecture's concentration process, self-regressive input, target selection, and instructional costs. These are listed below and were previously covered. An improved two-stage attention system is created by using a technique that promotes monotonicity and unimodality in the alignment matrix following the Tacotron2's material- and GPS-based attention system [3]. This modification is essential for enhancing stability during inference, especially when there are external controls. The model is exposed to both the final ground truth output value and the initial projected value during training using a two-pronged feedback technique (i.e., inference mode and instructor forcing). At the time of inference, the anticipated value is replicated. The model also incorporates the parameters needed to anticipate the 80-dim mel cepstral characteristics from a separately trained LPC-Net neural vo coder. These traits—which we refer to as "LPC features"—include 22-dim vectors with 20 cepstral coefficients, log f0, and f0 correlation for 22kHz signals. Instead of using post-net refinement, the mel task processes the anticipated Lcp elements using two put up-nets (one to enhance the area was found and one to enhance the pitch-related parameters).

$$L = \text{MSE}(\tilde{y}^M t, y^M t) + 0.8\text{MSE}(\hat{y}^L t, y^L t) + 0.4\text{MSE}(\tilde{y}^L t, y^L t) + 0.4MSE(\Delta\tilde{y}^L t, y^L t), \tag{3.1}$$

The modulo operator applies the starting time interval to the sequence, and MSE(,) represents the mean-squared error. To save space, we remove some information from this exposition and direct the reader to [5, 2] for more context and formulas.

3.2. Traditional Monitoring. This architecture depends on a Boolean indicator feature when labelled data is available. The audio signals' ground truth values are applied during training.

3.3. Without Supervision. The structure denoted by the letters A, C, D, as well as E, and so F, G, and H provide a way to increase the system's responsiveness during learning and to control the implementation of the prosodic rhythms at the time of inference using a configurable array of parameters (cf. the integer power of the controlled design) [6].

3.4. Mixed. Although parts D through I result from an unattended strategy, prosodic patterns may still be realized even in the presence of labelled data by working in conjunction with an explicit feature. We examine a "mixed" approach—specified by the entire framework Any–G—that combines controlled learning with technology to handle the circumstance without access to investigate this further [1].

3.5. Tiered Standard Mode Prosodic. The "layered standard pattern prosody" refers to an analytical model of language prosody, which is used to describe the organization of sound rhythm and phonological structure in language. This model typically includes multiple levels or levels to help understand different aspects of language rhythm. Following the motivation for a perceptually-interpretable, low-dimensional prosodic control mechanism discussed. We suggest an ordered collection of four prosodic regulators to condense information about a signal's length and pitch travel through linguistically significant and natural regions of the prosodic hierarchy. The strategy made it possible to manipulate common traits like general tempo. However, more control was needed to achieve the level of departure from long-term trends necessary to produce local emphatic concentration. These regulations apply to both domestic and international properties. They are a development of that strategy. Before we get there, let's define the following statistics [14].

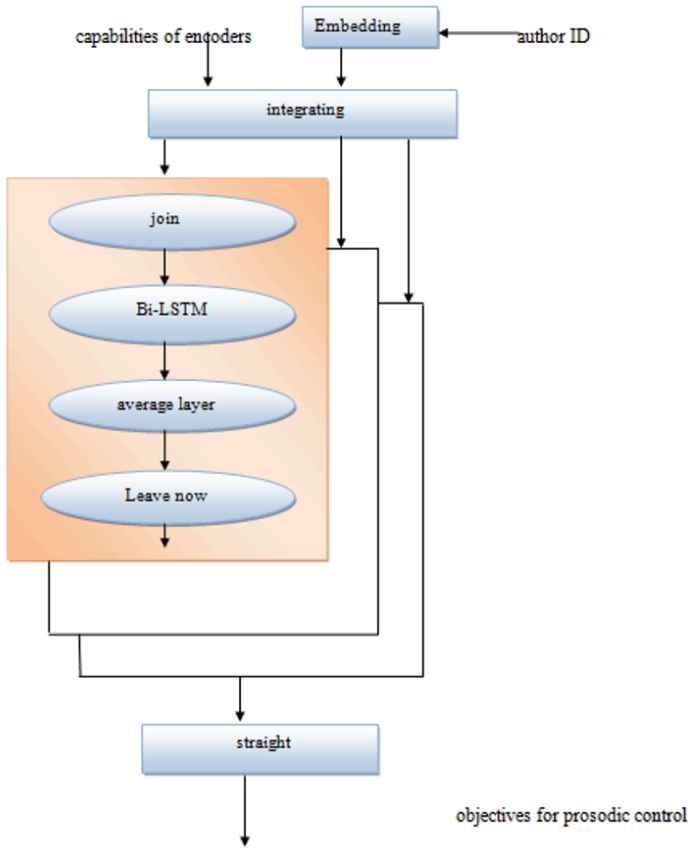

Fig. 3.1: Working flowchart of proposed Bi-LSTM for prosodic control

- S_{dur}: A record of the typical phone durations along a sentence (excluding quiet).
- S_{f0}: A "spread" of log-f0, which is the difference between its 95th and 5th percentiles, along a sentence.
- W_{dur}: A word-by-word log of the typical per-phone durations (discussed above).
- W_{f0}: A log-f0 "spread" along each word (as stated above).

$$PC = Norm_{\sigma}[S_{dur}, S_{f0}, W_{dur} - S_{dur}, W_{f0} - S_{f0}], \tag{3.2}$$

At the moment of inference, the prosodic-control subnet's predictions are adjusted to be roughly constant concerning the divination readings used to train the system. The evaluated systems' forecast distance between (known) phrase and term surround is stabilized using a mean pooling function. The design of the prosodic-control classifier. Models are trained using a speaker-embedding layer, whose output is fed into each tumbled interfere, utilizing a multi-speaker technique. We'll discuss how we create object sizes for each component of this architecture when we discuss its various elements [9]. Figure 3.1 describes the working flowchart of proposed Bi-LSTM for prosodic control.

4. Experimentation and Results. Three datasets from three native US English users who are employed as professionals made up the instructional stuff, which was broken down as follows:

- A canon from a male speaker (M1) with approximately 10.8K sentences.
- A corpus from the same male speaker (M1emp) with about 1K sentences containing multiple words with emphasis.

As part of the corpus M1emp, a speaker was taught to realise an emphatic prominence on the words that should receive the most emphasis inside each penalty. His realisations of prosody depart greatly from

broad focus prosody in terms of tempo, comparative diameter, stress height, and disjunction from the pertinent content. The sentences were meant to provoke certain limited-focus situations, such as contrast, disambiguation, etc.

Keep in mind that this sample is much smaller than the fundamental texts, and that only one speaker is included in the tagged data. On average, three emphatic words were present in each sentence in M1emp, and their overall frequency was roughly 23%. We are interested in comparing entirely unsupervised approaches that are feasible within the framework outlined in Section 2 with processes that use labelled data (where available) to determine the relative merits of each method.

To achieve that goal, take into account the following systems:

- Basis (NoEmph): A typical sentence-to-sentence communication system solely using worldwide prosodic limitations. The emphatic data is part of the learning batch (Demp), but no other focus-marking element is used.
- Basis (Sup): A basic system that uses Traditional Monitoring (as described in Section 2), world supervision, Demp training, and an apparent byte attribute expressing to determine the stress area.
- PC-Unsup: A fully unattended system with changeable prosodic control (as said in Section 2), where the prosody forecast and parts have been taught using Dbase.
- PC-Hybrid: a combination of models trained with Demp that gives precise Conditional accent signals and changing prosodic control, comparable to the Standard (Sup) system.

Comprehensive explanations and evaluation of the Base (NoEmph) layout with worldwide controllers may be found in [5]. However, in this instance, it acts as a reliable reference point for overall excellence to ensure that the alternative concepts support the benefits of naturalness provided by this technique [20]. LPC-Net is a vocoder model used in the field of speech synthesis and processing. It combines Linear Predictive Coding (LPC) and neural network technology to generate natural speech synthesis. Speaker training is typically used for voice modeling of vocoder models to simulate the speech features of different speakers. The model was chosen and tuned using the following steps. In order to scan the grid across architectures and track the held-out loss to gauge learning speed, 10% of the prosodic sub-networks training data were first withheld.In both instances, the speaker embedding had a size of 20. The remaining extreme parameters of the various combinations were then perceptually adjusted after this was taken care of.The beneficial boost rates that we choose are consistent with the empirical findings in the M1emp group and our theoretical hypotheses, which show that specialised items have longer speaking durations and more prominent tone accents. We found that the Base (Sup) system boosted our pitch incursions when adjusting the Hybrid systems because it recognised these tempo shifts rather well. After fine-tuning a single set of boosting parameters, we frequently find that it performs brilliantly, spanning a variety of words and dialects. Figure 4.1 defines the examples of the four prosodic controllers' phonetic trajectories for a two-sentence input.

5. Conclusion. We have developed and tested a method that enables more precise word syntax management to direct a comprehension of tight focus in the composite. The system is composed of consumer-driven regulations that follow the ideas we have articulated and supported. They offer a structure that divides different prosodic components (duration and pitch) so they can be altered separately. They are intuitive in the sense that changes to the way control is passed on to visual perception affect the outcome. They portray prosody in a flat manner. We have demonstrated that the method only requires additional data for simple word alignments. Different levels of monitoring can be accommodated with the necessary resources. Overall, deep machine learning has greatly improved the performance of intelligent speech recognition, enabling ASR systems to be widely used in daily life. However, there are still some challenges that need to be addressed, such as improving the robustness and accuracy of high-end end-to-end ASR systems to meet the needs of different application fields. This field is still constantly developing, and there will be more innovation and improvement in the future.

Acknowledgement. Supported by the university-level platform of Digital Creative Design Production-Education Integration Collaborative Innovation Center

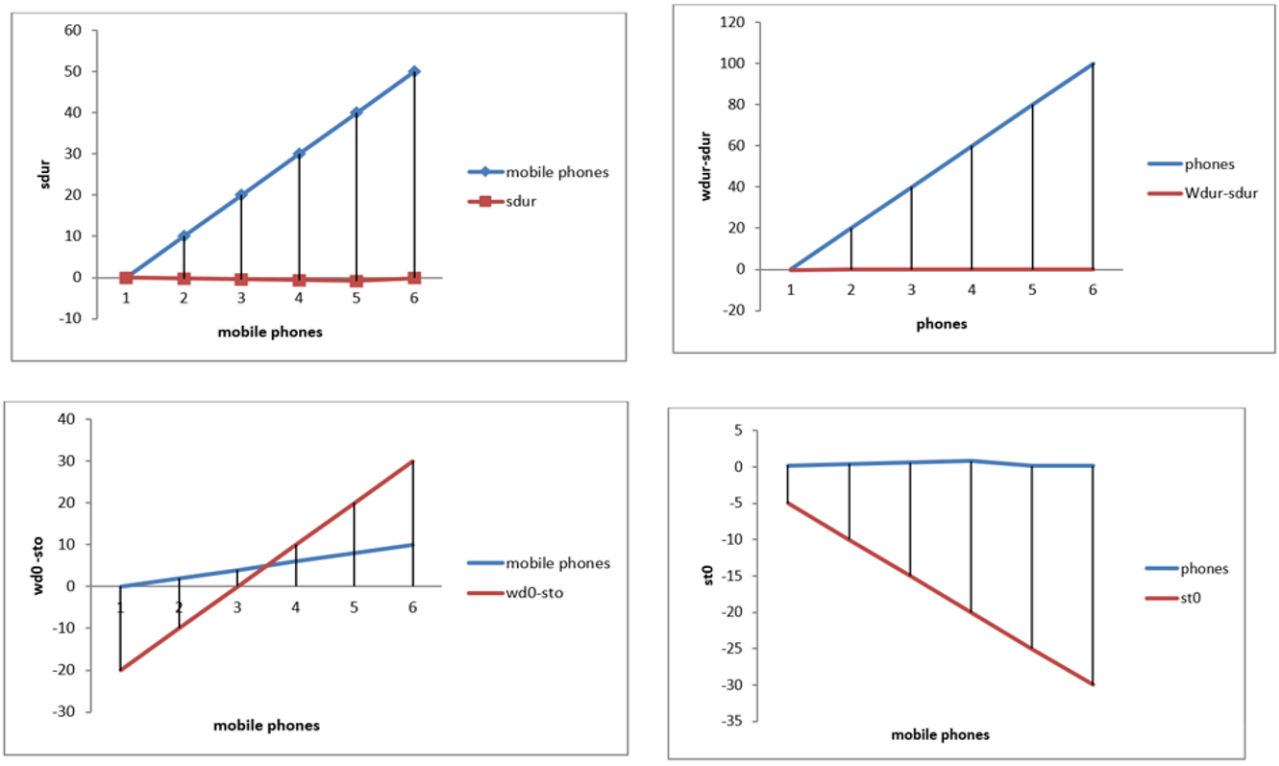

Fig. 4.1: Examples of the four prosodic controllers' phonetic trajectories for a two-sentence input

REFERENCES

[1] P. Ajay, B. Nagaraj, and J. Jaya, *Bi-level energy optimization model in smart integrated engineering systems using wsn*, Energy Reports, 8 (2022), pp. 2490–2495.

[2] P. Ajay, B. Nagaraj, R. A. Kumar, R. Huang, and P. Ananthi, *Unsupervised hyperspectral microscopic image segmentation using deep embedded clustering algorithm*, Scanning, 2022 (2022).

[3] J. L. Ba, J. R. Kiros, and G. E. Hinton, *Layer normalization*, arXiv preprint arXiv:1607.06450, (2016).

[4] R. Fernandez and B. Ramabhadran, *Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis*, in Proceedings of the 6th ISCA Workshop on Speech Synthesis, vol. 90, Bonn, Germany, 2007.

[5] W. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, *Hierarchical generative modeling for controllable speech synthesis*, in Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 2019, OpenReview.net.

[6] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in Proceedings of the 3rd International Conference on Learning Representations, Y. Bengio and Y. LeCun, eds., San Diego, CA, USA, 2015.

[7] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, *Fine-grained robust prosody transfer for single-speaker neural text-to-speech*, arXiv preprint arXiv:1907.02479, (2019).

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

[9] Y. Mass, S. Shechtman, M. Mordechay, R. Hoory, O. S. Shalom, G. Lev, and D. Konopnicki, *Word emphasis prediction for expressive text to speech*, in Proceedings of the 19th Annual Conference of the International Speech Communication Association, B. Yegnanarayana, ed., Hyderabad, India, 2018, ISCA, pp. 2868–2872.

[10] 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, 14 (2006), pp. 1099–1108.

[11] 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*, arXiv preprint arXiv:2006.04558, (2020).

[12] A. Sharma, A. Singla, N. Sharma, D. Gowda, et al., *Iot group key management using incremental gaussian mixture model*, in Proceedings of the 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC),

Coimbatore, India, 2022, IEEE, pp. 469–474.
[13] S. Shechtman, C. Rabinovitz, A. Sorin, Z. Kons, and R. Hoory, *Controllable sequence-to-sequence neural tts with lpcnet backend for real-time speech synthesis on cpu*, arXiv preprint arXiv:2002.10708, (2020).
[14] S. Shechtman and A. Sorin, *Sequence to sequence neural speech synthesis with prosody modification capabilities*, CoRR, abs/1909.10302 (2019).
[15] 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 Proceedings of the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), Calgary, AB, Canada, 2018, IEEE, pp. 4779–4783.
[16] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, *Modelling prominence and emphasis improves unit-selection synthesis*, in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2, 2007, pp. 1282–1285.
[17] 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), Barcelona, Spain, 2020, IEEE, pp. 6264–6268.
[18] J.-M. Valin and J. Skoglund, *LPCNet: Improving neural speech synthesis through linear prediction*, in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, IEEE, pp. 5891–5895.
[19] K. Yu, F. Mairesse, and S. Young, *Word-level emphasis modelling in hmm-based speech synthesis*, in Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, Texas, USA, 2010, IEEE, pp. 4238–4241.
[20] Y. Zhang, S. Pan, L. He, and Z. Ling, *Learning latent representations for style control and transfer in end-to-end speech synthesis*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, United Kingdom, 2019, IEEE, pp. 6945–6949.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 18, 2023
Accepted: Nov 25, 2023