PROSODY UNDER CONTROL: CONTROLLING PROSODY IN TEXT-TO-SPEECH SYNTHESIS BY ADJUSTMENTS IN LATENT REFERENCE SPACE

Juraj Šimko, Tuukka Törö, Martti Vainio, Antti Suni

University of Helsinki, Finland
firstname.secondname@helsinki.fi

ABSTRACT

We present a methodology for controlling prosodic characteristics of the output of an end-to-end TTS system based on direct adjustments of latent reference style conditioning vectors (style embeddings). The adjustments follow directions of maximal change calculated as gradients of regression fits of prosodic features against dimensions of the latent space. We also introduce a procedure for mitigating inter-dependencies among control in multiple prosodic dimensions based on an orthogonal projection. As the method does not rely on explicit prosodic annotations used for training, it is easy to implement, and can be applied to existing models without retraining. Our approach is evaluated on two single-speaker speech corpora that contain prosodically and stylistically rich material.

Index Terms: speech synthesis, style embeddings, prosodic features, explicit control, disentanglement

1. INTRODUCTION

As neural speech synthesis achieved human-level performance in reading aloud neutral sentences, the attention of the community has shifted towards methods of eliciting different speaking styles and rich prosodic variation ubiquitous in real human speech. Simply adding prosodic variety and different speaking styles to training data is not enough; the unaccounted variation present in the corpus is detrimental to the synthesis quality and leads to unpredictable and uncontrollable results.

In order to capture prosodic variation, utterance-level latent neural representations, *style embeddings*, are often used as an additional conditioning of the synthesis system [1, 2]. During training, the target acoustic representation identical to the systems output can be added as an additional input, and fed by a reference encoder through a heavily constrained bottleneck, before being concatenated with the textual input. The bottleneck style embeddings encode the recoverable (prosodic and stylistic) variation that is not inferrable from the text alone. Given an appropriate architecture, this approach has been shown to achieve a considerable degree of disentanglement of prosodic information from the text.

The style embedding latent space encodes the prosodic and stylistic characteristics of speech in a high-dimensional form that does not allow for a straightforward control of these characteristics in the synthesized output. Modifications to the encoder architecture [3], and learned pseudo-label approaches [4] have been applied to disentangle these complex representations or reduce the dimensionality, but robust control in terms of prosodic characteristics is still very much an open problem.

In this work, we present and evaluate a methodology for identifying the way in which relevant prosodic characteristics of speech are encoded in the latent embedding space and devise a mechanism for eliciting prosodic variation in a controlled way. The underlying assumption is that the latent space encodes the prosodic features in an entangled, but topologically coherent fashion. Broadly following the ideas presented in, e.g., [5, 6], we show that by following the gradients in the style embedding space, corresponding to systematic change in the investigated prosodic features of the encoded utterances, we can achieve a robust control of prosodic characteristics. We also present a method for “dissociating” the prosodic control, i.e., for minimizing the effects of elicited change in some prosodic features on other features of interest.

The proposed methodology is implemented using *reference prosody embeddings* trained within a Tacotron2 speech synthesis architecture as introduced in [2], and trained on two corpora of Finnish and English speech material (see Section 4).

2. PROSODIC FEATURES

Our approach is evaluated for four phonetic-prosodic characteristics of speech utterances derivable directly from speech signal: f_0 mean, f_0 standard deviation, spectral tilt, and speaking rate. Perceptually, these features correspond to overall pitch level, liveliness, voice quality, and tempo, respectively.

For utterances in the corpora as well as for synthesized utterances, the frequency related features were extracted by calling Praat [7] with the Parselmouth Python library [8]; speaking rate was approximated using text length as a proxy for the number of phonemic units. In more detail:

- f_0-MEAN and f_0-STD (standard deviation) were computed in semitones with the pitch floor at 75 Hz, ceiling at 400 Hz and a time step of 0.1;
- Spectral TILT was computed by analyzing the power spectrum to Long-Term Average Spectrum with a bandwidth of 100 Hz and computing its slope with the low band between 0 and 1 kHz and the high band between 1 and 4 kHz, using energy as the averaging method;
- Speaking RATE was calculated by dividing the orthographic length of the utterance by the total duration of the sound file.

3. PROSODIC CONTROL

Despite the complex nature of the underlying representation (see Fig. 1), the latent style embedding spaces have been shown to encode some relevant prosodic characteristics in a systematic way [5, 6, 9, 10, 11, 12]. In what follows we use a linear regression approach to devise a way to “extract” this systematicity from an embedding space produced by a trained encoder. The regression fits are used as an approximation of the relationship between the embedding space and the prosodic features of interest.

3.1. Relationship between the latent embeddings and prosodic features

Let Y be a set of n-dimensional style embedding vectors, obtained by running the training data through a trained reference encoder. In order to achieve comparable variance along all dimensions, we first Z-score normalize the vectors in Y by scaling. Let X be the resulting normalized version of Y with means equal 0 and standard deviations equal 1 along each dimension. Let *feat* be a (single dimensional) vector of phonetic-prosodic features of interest, calculated for the utterances encoded by X (e.g., an f_0-MEAN for each utterance).

A linear regression fit $feat \sim a_0 + a_1x_1 + \cdots + a_nx_n$ provides the best (in terms of RMS distance) linear approximation of the dependence of *feat* on the normalized embedding vectors. The regression coefficient vector $\mathbf{a} = (a_1, \ldots, a_n)$ is the gradient of this linear fit ($a_i \approx \frac{\partial feat}{\partial x_i}$); the vector provides the direction of the maximal slope of the fitted hyperplane, i.e., an estimate of the direction within the space X along which the feature *feat* exhibits the maximal increase.

The numerical values of the gradient vector $\mathbf{a}$ depend on the values and units of the phonetic-prosodic feature in *feat*. In order to “fit” the movement captured by vector $\mathbf{a}$ within the space X, the vector is linearly scaled so that the greatest absolute value along any dimensions equals 1 (by applying $\frac{\mathbf{a}}{\max |a_i|}$). As the standard deviation of the normalized embedding vectors is 1 along every dimension, this means the “movement” in the direction of the normalized vector $\mathbf{a}$ corresponds to maximally 1 standard deviation distance along any dimension in space X.

The normalized direction vector $\mathbf{a}$ is recast back to the original latent space Y by multiplying by the original standard deviation along every dimension. The resulting vector $\mathbf{b}$ provides an estimate of the direction in the reference prosody encoding space Y of the steepest positive change in the values of the feature of interest, and thus presumably allows for an explicit control of the synthesized speech in terms of the feature *feat* (see Section 4).

3.2. Dissociation by orthogonalization

The procedure described above provides an estimate of the direction of maximal change for a single given prosodic-phonetic feature. Given a possible entanglement of the way different features are encoded in the latent space Y (arising from potential correlations in the corpus), the process does not guarantee that the given direction does not considerably influence the other prosodic characteristics of synthesized speech. Systematically increasing mean f_0 might lead to an increase in, for example, speaking rate, if this type of relationship exist in the corpus. In order to attempt to dissociate the mutual inter-dependency of the controlled features we propose an approach based on orthogonal projection.

Let $\mathbf{a}_{feat_0}, \ldots, \mathbf{a}_{feat_m}$ be the direction vectors for $m+1$ features of interest, and let $\mathbf{F} = (\mathbf{a}_{feat_1}, \ldots, \mathbf{a}_{feat_m})$ be a matrix with columns corresponding to the feature vectors $\mathbf{a}_{feat_1}, \ldots, \mathbf{a}_{feat_m}$. Given an assumption of linear independence of these vectors, the matrix $\mathbf{F}$ is a basis of a hyperplane in the space X. The orthogonal projection $P\mathbf{a}_{feat_0} = \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{a}_{feat_0}$ yields the best approximation of the vector $\mathbf{a}_{feat_0}$ in the hyperplane $\mathbf{F}$. The vector $O\mathbf{a}_{feat_0} = \mathbf{a}_{feat_0} - P\mathbf{a}_{feat_0}$ is then the best approximation of the vector $\mathbf{a}_{feat_0}$ in the orthogonal complement of the hyperplane $\mathbf{F}$; it is orthogonal to all vectors in the space. In other words, the vector $O\mathbf{a}_{feat_0}$ depicts the best approximation of the direction (in the normalized space X) yielding maximal increase for the feature $feat_0$ and minimal (in terms of the linear regression fit) increase for all other features $feat_1, \ldots, feat_m$. In our subsequent analysis, we will use both these orthogonalized and non-orthogonalized direction vectors in order to dissociate the effects of prosodic control proposed here.

4. EVALUATION

In the present work we used a Tacotron 2 architecture [13], extended with a reference encoder. The implementation of the reference encoder matches the architecture described in [2]; the mel-spectrogram is processed and downsampled with six 2D convolutional layers, followed by a gated recurrent unit layer (GRU), and the final output of the GRU is taken as the style embedding. This 128 dimensional embedding is then concatenated with each character embedding of the utterance, forming the input to the decoder.

In order to evaluate the proposed prosodic control method, we trained two voices. The first voice was trained on the commonly used LJSpeech dataset [14], 24 hours of relatively lively narrative style read by an American English female speaker. The second voice was trained on a new, yet to be published Finnish speech corpus, FinSyn, of two professional female speakers, with duration of approximately 30 hours each. The corpus was designed with “ordinary” speaking style variation in mind, containing mostly continuous read speech of texts with varying degrees of formality, as well as semi-spontaneous and spontaneous speech. A 22 hour subset from a single speaker was used for training.

Both voices were trained on orthographic text, with numbers and abbreviations expanded. For the Finnish voice, we applied a version of curriculum learning, training the system first on a more formal utterances of isolated sentences until diagonal attention was achieved, after which the rest of the data was added to the training set. Both voices were trained from scratch for 200,000 steps with a batch size of 16, on a single Nvidia V100 GPU.

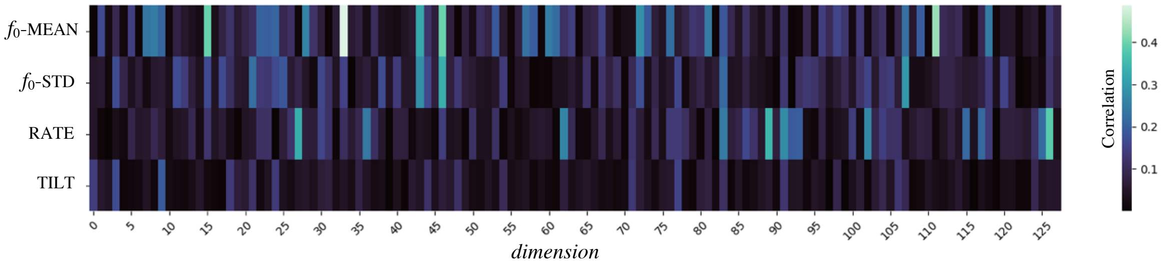

Figure 1: Correlations between the prosodic features and dimensions of the style embedding reference space.

Table 1: Slopes and r^2-values (in brackets) of the fits with scaling factor of the directional vector for the controlled features (in columns) as predictors and the feature values extracted from the synthesized utterances as dependent variables (in rows).

	non-orthogonal direction vectors				*orthogonal direction vectors*			
	f_0-MEAN	f_0-STD	TILT	RATE	f_0-MEAN	f_0-STD	TILT	RATE
Finnish corpus								
f_0-MEAN	**1.43** (*0.92*)	0.88 (*0.71*)	-0.15 (*0.25*)	0.18 (*0.45*)	**1.26** (*0.95*)	0.37 (*0.47*)	-0.20 (*0.37*)	0.18 (*0.30*)
f_0-STD	0.11 (*0.12*)	**0.65** (*0.90*)	-0.13 (*0.38*)	-0.05 (*0.09*)	0.10 (*0.12*)	**0.72** (*0.94*)	0.00 (*-0.01*)	0.17 (*0.54*)
TILT	-0.35 (*0.31*)	-0.53 (*0.50*)	**1.29** (*0.90*)	-0.05 (*0.00*)	-0.28 (*0.29*)	-0.36 (*0.33*)	**1.24** (*0.91*)	-0.25 (*0.27*)
RATE	0.05 (*0.03*)	-0.04 (*0.01*)	0.12 (*0.17*)	**0.41** (*0.66*)	0.05 (*0.02*)	0.01 (*-0.01*)	0.09 (*0.09*)	**0.37** (*0.62*)
English corpus								
f_0-MEAN	**1.53** (*0.97*)	0.38 (*0.5*)	-0.12 (*0.13*)	0.2 (*0.36*)	**1.47** (*0.96*)	0.35 (*0.45*)	-0.11 (*0.08*)	0.09 (*0.11*)
f_0-STD	0.07 (*0.07*)	**0.64** (*0.93*)	0.05 (*0.06*)	-0.06 (*0.14*)	0.01 (*0.00*)	**0.61** (*0.89*)	0.01 (*0.00*)	0.01 (*0.00*)
TILT	-0.12 (*0.02*)	0.13 (*0.03*)	**2.01** (*0.89*)	0.14 (*0.04*)	-0.1 (*0.01*)	-0.10 (*0.02*)	**2.00** (*0.89*)	0.24 (*0.11*)
RATE	0.07 (*0.03*)	-0.08 (*0.06*)	-0.01 (*-0.01*)	**0.36** (*0.48*)	0.04 (*0.01*)	-0.07 (*0.04*)	0.03 (*0.00*)	**0.30** (*0.39*)

4.1. Evaluation of prosodic control

For each of the two trained models (the Finnish and English voice), the style embedding vectors were extracted for each utterance in the training corpus using the reference encoder. Also, the values of the prosodic features were calculated for the same utterances, as described in Section 2. Fig. 1 shows the correlations between the individual dimensions of the embedding space and the prosodic features. As seen on the figure, each prosodic feature correlates with multiple dimensions of the embedding space, and several dimensions relatively strongly correlate with multiple features.

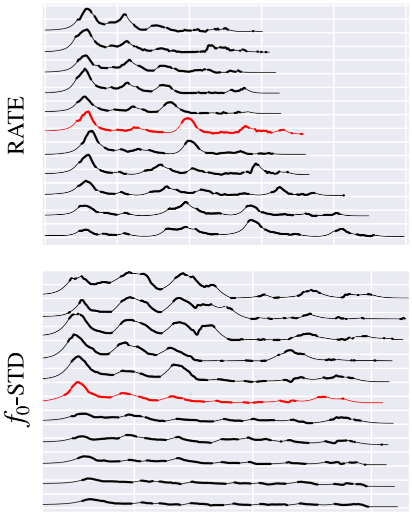

Figure 2: f_0 trajectories of utterances generated with the Finnish system, controlling speech rate and f_0 standard deviation, zero modification in red

The efficacy of the method for disentangling this complex relationship and for providing a control over the prosodic characteristics was evaluated in the following way. Both non-orthogonal and orthogonal versions of the directional vectors for the investigated features (vectors **b** from Section 3) were calculated using the style embedding spaces and prosodic features. Each direction vector was orthogonalized against the hyperspace containing the direction vectors of all three remaining features. The direction vectors were then scaled by the factors corresponding to integers between -5 and 5 (i.e., up to 5 standard deviations of the original distribution), and added to the mean vector of the entire embedding space.

For each scaled version, ten Harvard sentences/their Finnish translations were synthesized and the prosodic characteristics of the synthesized sentences were extracted (see https://tuukkaot.github.io/PhoneticFeatureTTS). Fig. 2 shows the resulting utterances obtained by scaling the direction vector for f_0-STD and RATE feature.

Fig. 3 shows the distributions of the prosodic characteristics of the test sentences (in rows) as a function of the directional vector scaling (in columns) for the Finnish corpus (the corresponding plot for the English corpus shows very similar behaviour). The plots on the diagonal, depicting the behaviour in terms of the controlled feature, show a strong albeit not always linear relationship between the scale and the controlled prosodic characteristics; increasing the scale leads to a systematic increase of the given feature. The plots off the diagonal in Fig. 3 generally show considerably weaker influence of the control of on the non-controlled prosodic characteristics.

In order to quantify these observations, a linear regression model with feature values of the synthesized utterances as the dependent variable and scaling factors as predictors were fitted for every combination captured in Fig. 3; the fits are shown as the lines in the figure. The slopes of these fits and the adjusted r^2-values of the models (depicting the quality fit) are listed in Table 1. The table summarises the models for both corpora and for both non-orthogonal and orthogonalized the direction vectors.

The quality of the fits shows that the control exerts reliable and strong influence on the controlled features. The r^2s and the slopes of the fits are greater on the diagonals (in bold) than off diagonals, i.e, we get better quality of fits and stronger relations for the controlled features than for the non-controlled ones (as the slopes depend on the units of the dependent variables, only the values in the same row are directly comparable).

Regarding the effects of orthogonalization, in the majority of cases both slopes and r^2-values off diagonals are lower for the orthogonalized directional vectors then for the non-orthogonalized ones. For example, the strong influence of f_0-STD control on the f_0-MEAN feature for the Finnish corpus, clearly visible in Fig. 3, gets substantially attenuated: the slope decreases from 0.88 to 0.37, r^2-value from 0.88 to 0.37.

For some feature combinations, primarily related to speech rate control, the orthogonalization did not work as envisaged; e.g., the effects of RATE control on TILT feature actually *increased* for both corpora.

5. DISCUSSION

The presented method of controlling prosodic features in TTS by directly manipulating style embedding space elicited considerable variation in the controlled features. Because the method constructs (approximates) the reference vectors in the existing trained latent space, the prosodic characteristics are not controlled explicitly, but the synthesis rather relies on—and replicates—the relevant variance in the training material. For example, the observed speech rate variation is augmented by appropriately inserted/removed silent pauses rather then by a uniform stretching of the synthesised output (see Fig.2). The synthesis quality is thus not seriously compromised even for the relatively high values of the scaling factors.

While we have evaluated the prosodic control only for a relatively small set of prosodic characteristics, the methodology (unlike the explicit prosodic control systems, e.g., [15, 16]) allows for a *post hoc* inclusion of any quantifiable controlled features in a straightforward way by simply calculating and scaling the appropriate direction vectors (see, e.g., [5, 6] using similar methodology). Also, we have chosen a relatively “greedy” approach to the orthogonalization of directional vectors by enforcing their disentanglement from all other investigated features. This might not be necessary in practice when the user of the TTS system might require independent control of a smaller combinations of prosodic characteristics (e.g., tempo from pitch level but not from voice quality). The proposed method accounts for this type of flexibility.

The method can be easily combined with other approaches to style control. While we added the scaled direction vectors to the mean style embedding of the corpus (equivalent to an “average” rendition), this starting point could be shifted to any point within the embedding space, for example, to a style embedding representing a particular style identified in the training material (by using appropriate explicit or learned labels).

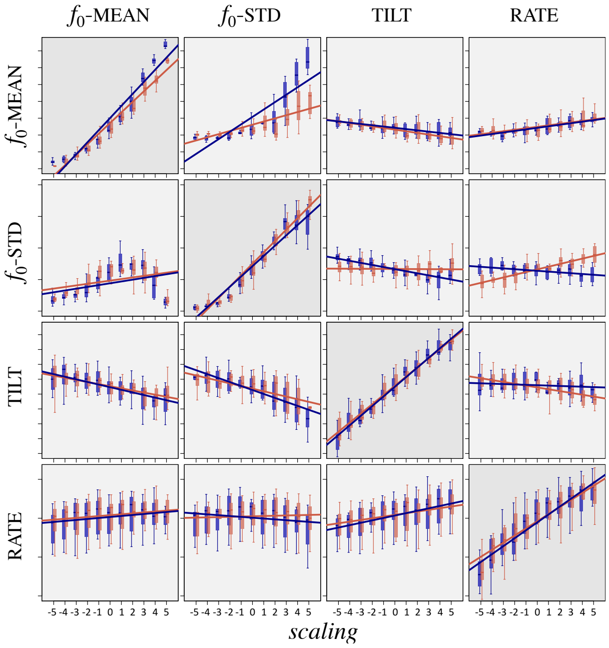

Figure 3: The distributions of the prosodic characteristics of the test sentences as a function of the directional vector scaling for the Finnish corpus. The columns correspond to the controlled features, the rows to the consequences of the control. The non-orthogonalized control in blue, orthogonalized in red.

Comparing the performance for the four tested features shows relatively less reliability of speech rate control compared to the other three features. The greater variability in RATE feature in the synthesized material (reflected by the lower r^2-values) is presumably a consequence of using the number of orthographic characters for calculating the RATE feature, and the complex relationship between the letters and durations of speech sounds. Also, in order to simplify the process, we used the duration of the sound file rather than duration of the actual utterance for the RATE feature calculation; leading and training silences (and also the pauses within the utterance) might contribute to the variance.

While the corpora used here were not explicitly designed with a vast range of variation along the controlled characteristics in mind, they do contain prosodically rich material. This richness does not compromise the synthesis quality produced by the used TTS system, but undoubtedly contributes to the degree of controllability. Design and collection of corpora containing prosodically varied material (rather than containing carefully crafted sentences and steady articulation) is one the necessary requirements for achieving high quality, and prosodically rich and controllable synthesis.

6. REFERENCES

[1] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4693–4702.

[3] G. E. Henter, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *CoRR*, vol. abs/1807.11470, 2018. [Online]. Available: http://arxiv.org/abs/1807.11470

[4] 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.

[5] N. Tits, F. Wang, K. E. Haddad, V. Pagel, and T. Dutoit, "Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis," *arXiv preprint arXiv:1903.11570*, 2019.

[6] N. Tits, K. El Haddad, and T. Dutoit, "Analysis and assessment of controllability of an expressive deep learning-based tts system," in *Informatics*, vol. 8, no. 4. Multidisciplinary Digital Publishing Institute, 2021, p. 84.

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

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

[9] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Proc. Interspeech*, 2017.

[10] A. Suni, M. Wlodarczak, M. Vainio, and J. Šimko, "Comparative analysis of prosodic characteristics using wavenet embeddings," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019) Crossroads of Speech and Language*. ISCA, 2019.

[11] J. Šimko, M. Vainio, A. Suni *et al.*, "Analysis of speech prosody using wavenet embeddings: The lombard effect," in *Proceedings of 10th International Conference on Speech Prosody 2020, Tokyo, Japan*. ISCA, 2020.

[12] K. Hiovain, A. Suni, S. Kakouros, and J. Šimko, "Comparative analysis of majority language influence on north sámi prosody using wavenet-based modeling," *Language and Speech*, p. 0023830920983591, 2020.

[13] 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.

[14] K. Ito, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[15] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[16] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *arXiv preprint arXiv:2009.06775*, 2020.