EVALUATING THE IMPACT OF PROSODY FEATURE NORMALIZATION ON THE CONTROLLABILITY OF PITCH IN SPEECH SYNTHESIS

Judith Bauer[1], *Frank Zalkow*[1], *Meinard Müller*[1,2], *Christian Dittmar*[1]

[1]*Fraunhofer IIS, Erlangen, Germany,* [2]*International Audio Laboratories Erlangen, Germany*
judith.bauer@iis.fraunhofer.de

Abstract: Recent neural text-to-speech (TTS) models are able to synthesize highly natural speech signals using deep learning techniques. In practical applications, it can be desirable to have explicit control over the prosody (speech rate, fundamental frequency, and energy) of the synthesized speech. Such controllability can be achieved by adding prosody prediction modules, whose main purpose is to estimate plausible prosody features for each phoneme in the text input. This explicit modeling also allows for changing prosody features at inference time, consequently enabling the adjustment of the prosody in the synthesized audio. In this paper, we evaluate to which extent deliberate manipulation of such prosody features is reflected in the resulting speech audio. We focus particularly on changing the pitch (i.e., fundamental frequency) while applying different normalization strategies.

1 Introduction

Todays state-of-the-art text-to-speech (TTS) systems usually consist of two main components: Textual input, given as phoneme sequence, is first processed by the so-called *acoustic model*, whose main task is to predict mel spectrograms, which implicitly encode the information about how a text is uttered (including, e.g., duration of phonemes, speech melody). Since a text can be spoken in many different ways, even by the same speaker, the mapping from textual input to mel spectrograms is a one-to-many task. The second component is the *neural vocoder* which takes mel spectrograms and converts them into time-domain speech waveforms. This mapping is more unique since the mel spectrogram already dictates how a plausible speech waveform should be structured.

For many practical applications, it can be desirable to have more fine-grained control over the mel spectrogram prediction process of the acoustic model. A natural choice is to modify the prosodic features which represent interpretable speech characteristics, such as phoneme durations (speech rate), fundamental frequency (referred to as pitch in the following), and energy (loudness). A successful approach to explicitly model these features is the extension of the acoustic model by prosody predictors, which are trained jointly with the acoustic model. Training such an extended acoustic model requires the provision of ground-truth prosody features which can be extracted from the speech samples in the training dataset using standard signal processing methods. This additional information eases the one-to-many mapping task and leads to a disentangled internal representation of prosody and voice timbre. During inference, where no ground-truth prosody features are available, the previously trained prosody predictors take over the task to provide prosody information.

In the literature, there have been several proposals of acoustic model architectures as described above. FastSpeech2 [1] and FastPitch [2] were introduced in parallel, extending the popular FastSpeech [3] architecture with predictors for pitch and energy. FastTacotron [4] adds

 DOI: 10.35096/othr/pub-7097, URN: urn:nbn:de:bvb:898-opus4-70976

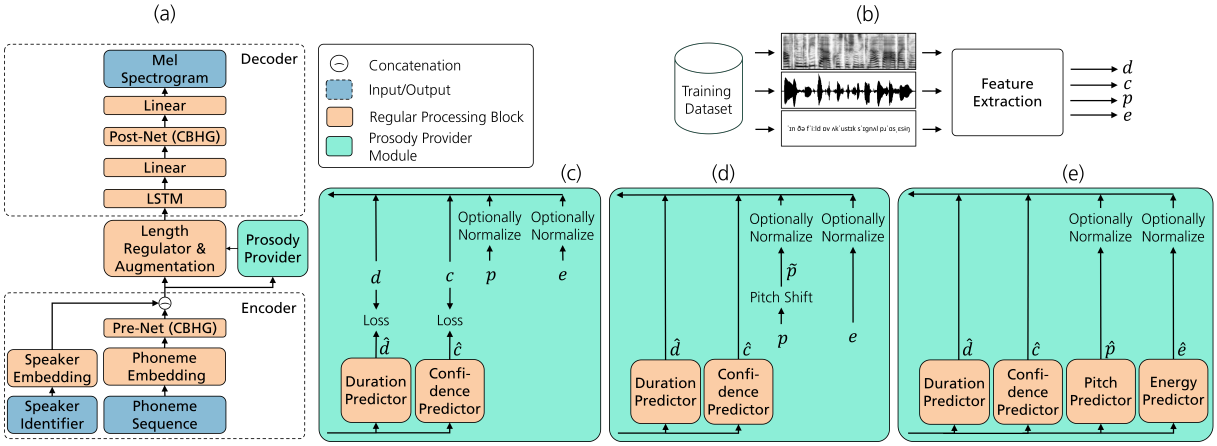

Figure 1 – Model overview with training and inference settings used in this paper. Subfigure (a) depicts the architecture of our acoustic model based on a modified ForwardTacotron. For details, see Sec. 2.2. Subfigure (b) illustrates the extraction of ground-truth prosody information from our training dataset, details see Sec. 2.1. The acoustic model can be trained circumventing some of the prosody predictors. Usage of ground-truth pitch and energy features is illustrated in (c) for training and (d) for inference. Subfigure (e) shows inference with the full set of prosody predictors.

pitch and energy predictors to the ForwardTacotron [5] architecture. The main difference between these two model families is the usage of feed-forward Transformer blocks (FastSpeech, FastSpeech2, FastPitch) vs. recurrent units (ForwardTacotron, FastTacotron).

To the best of our knowledge, only few publications address in detail the evaluation of the prosody predictors' effect on the actual prosodic features apparent in the synthesized speech. One such example is the work of Mohan et al. [6], who propose a modification of Tacotron2 [7] with pitch, energy, and duration predictors. The authors conduct an objective evaluation of the prosody features' entanglement by examining the effect of modifying one feature while keeping the others fixed. This evaluation shows that modifications of the prosody features correlate with the corresponding properties in the synthesized speech and that different prosody features can be changed relatively independently of each other. However, their evaluation does not report how accurately the desired modification is reflected in the synthesized speech audio.

In this paper, we try to address open questions from the papers discussed above: **(1) Normalization:** In previous work, the prosody features are often normalized. In our experiments, we use a multi-speaker training set and seek to find the optimal strategy with respect to feature normalization. We investigate three variants: leaving the prosody features unnormalized, normalizing them using statistics (mean and standard deviation) derived from the complete dataset, or normalizing them using statistics derived for each speaker individually. As a main contribution, we report how the normalization affects accuracy and pitch range of the synthesized speech. Therefore, our secondary contributions are: **(2a) Accuracy:** We examine how much control over the exact prosody characteristics in the synthesized audio samples is possible. In particular, we focus on the control over pitch values and investigate how accurately the specified pitch is reflected in the synthesized audio samples. This is relevant, for example, when giving a user control over prosody features, to ensure that the specified and the actual pitch values correspond as accurately as possible. **(2b) Range of pitch control:** We also investigate the influence of the speaker-dependent pitch distribution in the training data on the possible extent of pitch modification. We expect that specifying pitch values farther away from the average pitch of a speaker leads to lower accuracies.

2 Method

2.1 Data

For training the models in our experiments, we use an English dataset consisting of four speakers: "female1" [8] (5.44 h), "male1" [9] (5.43 h), "female2" (2.22 h), and "male2" (2.14 h), where the latter two are proprietary corpora recorded with professional voice actors. The dataset comprises pairs of weakly aligned phoneme annotations and speech recordings. The recordings are downsampled to a sampling frequency of 22 050 Hz before computing mel spectrograms (with 80 bands, 256 samples hop size, 1024 samples block size). As indicated in Fig. 1b, we extract ground-truth prosody features for each speech sample in our dataset. As a pre-requisite, the weak correspondence between the phoneme transcriptions and the corresponding mel spectrograms needs to be refined into a phoneme-wise alignment. More precisely, this alignment specifies the number of mel frames that are associated to each phoneme, and is henceforth denoted as d. To this end, we use an aligner model [10] that temporally aligns phoneme sequences to mel spectrograms. In addition to that, we extract frame-wise prosody features. These comprise energy as the L2-norm of the mel spectrogram frames, pitch in Hertz, and voicing confidence, which can be interpreted as saliency of the pitch estimates. Pitch and voicing confidence are extracted using CREPE [11]. Subsequently, the alignment information in d is re-used to summarize frame-wise prosody features to the phoneme level, yielding phoneme-wise energy e, pitch p, and confidence c. As proposed in previous works [2, 4], we thereby discard pitch estimates below a certain voicing confidence threshold.

2.2 Model

We use an acoustic model based on ForwardTacotron [5]. The model was extended with prosody predictors for phoneme duration, pitch, energy, and voicing confidence, as described by Zalkow et al. [12]. An overview of the architecture is shown in Fig. 1a. On a high level, the acoustic model consists of an encoder part where phoneme sequences and speaker identifiers are transformed into internal representations. Note that the sequence length stays the same as the number of phonemes in the phoneme sequence throughout all layers of the encoder. It is therefore much lower than the corresponding number of target mel spectrogram frames. The middle part serves three purposes: First, in the "prosody provider," a bank of prosody predictors estimates duration $\hat{d}$, pitch $\hat{p}$, energy $\hat{e}$, and voicing confidence $\hat{c}$ per phoneme based on the encoder output. Second, the internal feature representation is augmented with these estimates. Third, the so-called "length regulator" resamples the temporal axis of the internal feature representation in a non-equidistant fashion to the number of mel spectrogram frames using the durations from the prosody provider. The final building block of the acoustic model is the decoder, which predicts the mel spectrogram from the augmented internal representation.

As mentioned before, the prosody predictors are trained jointly with the encoder and decoder of the acoustic model. To that end, they receive the output of the encoder and learn to predict their respective prosody values $\hat{d}$, $\hat{c}$, $\hat{p}$, $\hat{e}$ by minimizing the L1-Loss to the ground-truth features d, c, p, e. It is important to note that the same ground-truth information is also used during training in a teacher-forcing paradigm. That means both the length regulator and the decoder are exposed to ground-truth prosody features throughout the entire training process. This helps to stabilize the training of the overall model since it can be expected that the prosody predictors perform poorly early in the training process.

At inference time, one has usually no access to ground-truth prosody features and has to rely on the prosody predictors as shown in Fig. 1e. However, for our experiments, we directly provide specific pitch and energy features which we can deliberately manipulate as shown in

Experiment	Feature Normalization Method			Augmentation Method		Genders in Training Data	
	orig	normGlob	normInd	concatenation	addition	female	male
a	✓			✓		✓	✓
b		✓		✓		✓	✓
c			✓	✓		✓	✓
d	✓				✓	✓	✓
e		✓			✓	✓	✓
f	✓			✓		✓	
g	✓			✓			✓

Table 1 – Overview of models with different feature normalization methods, augmentation methods, and genders in the training data.

Fig. 1d. In our evaluation, we focus on shifting the pitch (see Sec. 2.3.2) before passing it to the decoder. Since, in our experimental setup, we use only ground-truth pitch and energy values (both during training and inference), we remove the respective predictors from the acoustic model, as shown in Fig. 1c. This will also influence the learning of the internal feature representation in the encoder. Obviously, such a models is not usable for inference, but allows us to study the encoder and decoder independently from the prosody predictors.

As a neural vocoder, we use a pre-trained StyleMelGAN [13], which is fixed throughout all experiments of this paper. Since the voices from our training dataset were included in the vocoder training, we can expect good synthesis quality.

2.3 Experimental Setup

2.3.1 Feature Normalization

As shown in Fig. 1a, the encoder output of the acoustic model is augmented with the output of the prosody provider (i.e., pitch, energy, confidence, and duration information). Recall that the prosody predictor outputs are replaced by ground-truth prosody features d, c, p, e (see Fig. 1c) in order to realize teacher-forcing during training. In contrast, $\hat{d}$ and $\hat{c}$ are used instead of d and c during inference, see Fig. 1d. Unlike for duration and confidence, we always use ground-truth pitch and energy in the experiments of this paper, since we are interested in investigating the effect of feature normalization with regard to these features. Normalization in our case is realized by subtracting a mean value and dividing by a standard deviation. Since we operate in a multi-speaker setting, we can either base these statistics on the complete dataset (yielding global normalization) or on a per-speaker basis (yielding individual normalization). Regarding "no normalization" with features remaining at their original values as additional option, we have three normalization options, which are also shown in Tab. 1: no normalization (referred to as "orig"), global normalization (referred to as "normGlob"), and individual normalization (referred to as "normInd").

2.3.2 Pitch Shifting

We evaluate the success of the normalization approaches described above using 20 held-out pairs of phoneme sequences and speech recordings per speaker. To cover a pitch range which goes beyond the pitch distribution observable in the training data, we deliberately apply pitch shifts to the ground-truth pitch p from -1 octave to $+1$ octave in semitone steps. In the following, we refer to those pitch values as "expected pitch" and denote them for brevity by $\tilde{p}$. Then, we use our trained TTS system to synthesize speech with the given expected pitch. Since we

use 25 different pitch shifts, this strategy yields 25 different synthetic speech samples per test utterance, totaling to $\#speakers \cdot \#sentences \cdot \#shifts = 4 \cdot 20 \cdot 25 = 2000$ audio samples per experiment. The pitch of the synthesized speech samples is extracted using CREPE, and again summarized to the phoneme level. We refer to this as the "actual pitch", denoted by $\bar{p}$. Ideally, $\bar{p}$ should closely follow $\tilde{p}$, but, in practice, there are differences, which we quantify as pitch accuracy and discuss in Sec. 3. Obviously, this is a purely objective evaluation strategy, leaving subjective assessment of speech naturalness for future research.

2.4 Further Experiments

To receive some insights into how the pitch is used by the acoustic model, we conduct further experiments:

Comparison of addition and concatenation: In previous works, augmentation of the encoder output with prosody features is often realized by adding prosody values to the encoder output. This requires to process them by an additional layer to match the feature dimensions. We conduct experiments with concatenated, as well as added features, as shown in Tab. 1.

Single-gender models: We evaluate the pitch accuracy for each speaker separately. To verify our assumption that more diverse training data helps generalization, we train a model with only female speakers and a model with only male speakers (see Tab. 1: exp. f, g) and compare them to a model trained with all four speakers.

Arbitrary scaling of pitch: We investigate the principal influence of the range of values covered by the pitch features, regardless of plausible normalization. To this end, we arbitrarily upscale pitch features (when normalized, as in Tab. 1: exp. b, c) or arbitrarily downscale pitch features (when unnormalized, as in Tab. 1: exp. a).

3 Results

3.1 Evaluation of General Accuracy

We evaluate the experiments shown in Tab. 1, including the methods explained above for prosody feature normalization, the methods for augmenting the encoder output, and the different training sets. These setups are compared with regards to their ability to retain the expected pitch values in the synthesized audio samples. Therefore, we condition the models with the expected pitch $\tilde{p}$ and derive the actual pitch $\bar{p}$ as described in Sec. 2.3.2. To measure the distance between $\tilde{p}$ and $\bar{p}$, we compute the pitch accuracy as mean squared difference in semitones. The results are shown in Tab. 2. Experiments a–c entail the different normalization methods, indicating that the usage of unnormalized ("orig") prosody features yields the lowest error and per-speaker normalized ("normInd") prosody features yield the highest error in an otherwise fixed setup. The low accuracy associated with the per-speaker normalization can be expected, since in this scenario, the features convey less information about the absolute pitch to the decoder. With regards to the method for augmenting the encoder output with the values from the prosody provider, experiments d and e in Tab. 1 show that augmentation by addition leads to higher errors than augmentation by concatenation when tested in a setting without normalization ("orig"). When considering global normalization "normGlob," augmentation by addition leads to similar (slightly lower) errors compared to augmentation by concatenation. In experiments f and g, we train models on the data of only female (experiment f) or only male speakers (experiment g). Apart from the training data, the setup is similar to experiment a. For all speakers, the accuracy of the single-gender models is inferior to the accuracy of the model trained with the complete dataset, indicating that a more diverse training dataset is beneficial for the pitch accuracy. Furthermore, we modify experiments a and b by arbitrary downscaling

Experiment	female 1	female 2	male 1	male 2	Average
a (orig, concat)	**5.04** (±2.04)	**4.74** (±1.94)	**8.22** (±2.57)	**7.60** (±2.47)	**6.40**
b (normGlob, concat)	9.73 (±2.70)	10.15 (±2.62)	8.65 (±2.61)	8.18 (±2.54)	9.18
c (normInd, concat)	13.31 (±3.16)	14.25 (±3.16)	9.72 (±2.79)	12.77 (±3.21)	12.51
d (orig, add)	15.91 (±2.93)	15.77 (±2.83)	10.80(±2.65)	15.31 (±3.27)	14.45
e (normGlob, add)	**4.30** (±1.92)	**4.68** (±1.95)	**8.17** (±2.57)	**6.88** (±2.37)	**6.01**
f (orig, concat, only Female)	9.70 (±2.75)	9.17 (±2.63)			9.44
g (orig, concat, only Male)			10.30 (±2.89)	13.68 (±3.27)	11.99
a with arbitrary downscaling	11.47 (±2.99)	11.98 (±2.86)	8.09 (±2.51)	8.66 (±2.60)	10.05
b with arbitrary upscaling	3.26 (±1.65)	3.44 (±1.67)	9.44 (±2.80)	9.06 (±2.72)	6.30

Table 2 – Results of the experiments which were described in Tab. 1. The results are given as mean squared error and standard deviation of the distance between expeced pitch $\tilde{p}$ and actual pitch $\bar{p}$ in semitones. For details of the experiments using arbitrarily downscaled/upscaled pitch features, see Sec. 2.4.

and upscaling of the prosody features, respectively, as described in Sec. 2.4. The downscaling operation leads to a lower accuracy compared to the original experiment a, while the upscaling operation improves the accuracy compared to the original experiment b. Therefore, it can be assumed that features in a higher value range are beneficial for the accuracy.

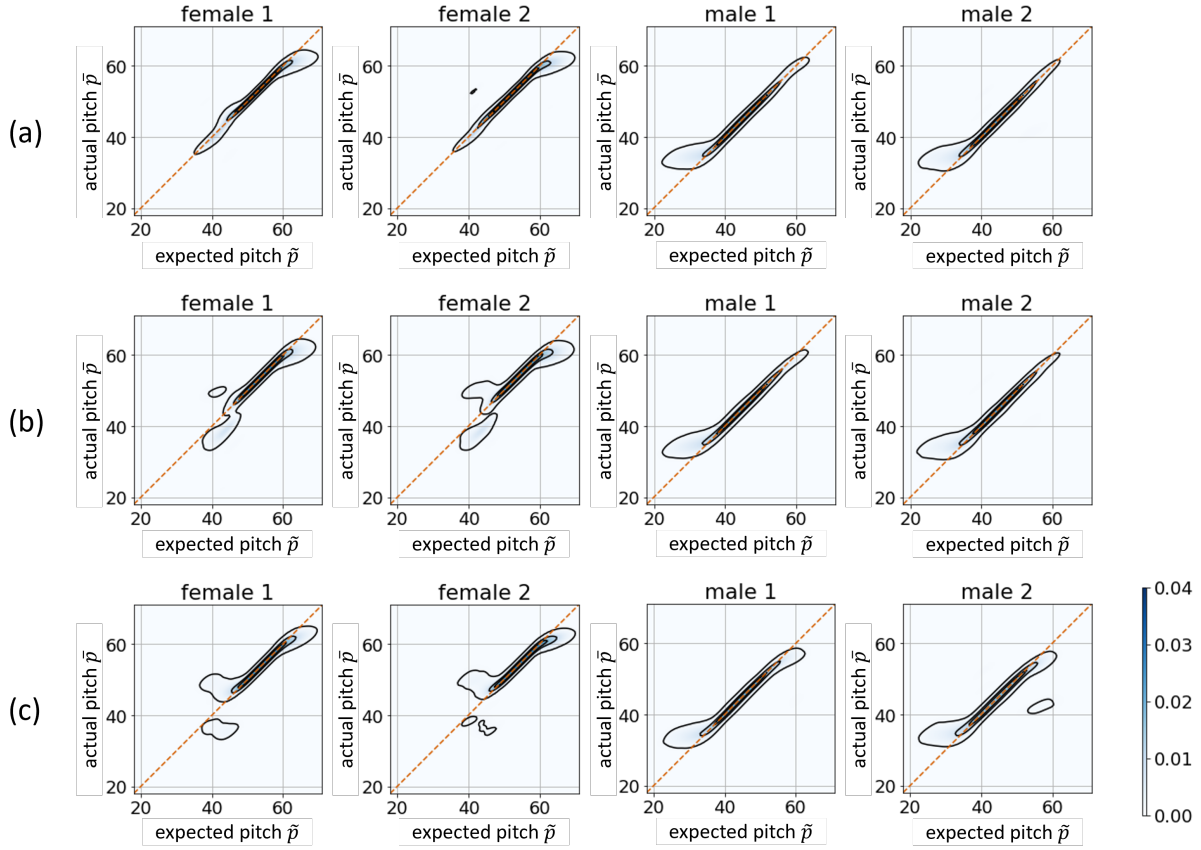

Figure 2 – Comparison of expected pitch $\tilde{p}$ and actual pitch $\bar{p}$ for (a) a model with unnormalized ("orig") pitch features (corresponding to experiment a in Tab. 1), (b) a model with globally normalized ("normGlob") pitch features (corresponding to experiment b in Tab. 1) and (c) two models with unnormalized ("orig") features which are trained on only female or only male speakers respectively (corresponding to experiments f and g in Tab. 1). The pitch is given in semitones with a base frequency of 10 Hz. Ideally, $\tilde{p}$ is similar to $\bar{p}$ (as indicated by the orange diagonal lines). The figure depicts the distributions of $\tilde{p}$ and $\bar{p}$, estimated using a kernel density estimate.

3.2 Evaluation of Accuracy Depending on Pitch Range

We intend to gain more insights into the relationship between the expected pitch features $\tilde{p}$ (i.e., features which are passed to the model) and the actual pitch values $\bar{p}$ (i.e., pitch features derived from the synthesized audio samples). To this end, we select four experiments from Tab. 1 for further investigation: Experiment a, which uses unnormalized ("orig") prosody features and corresponds to a high pitch accuracy as shown in Tab. 2, experiment b, which relies on globally normalized ("normGlob") features, experiment f, which reduces the training dataset to only female speakers, and experiment g with a training dataset including only male speakers. For these experiments, we plot the distribution of $(\tilde{p}, \bar{p})$-pairs as kernel density estimate plots in Fig. 2. Optimally, the expected and actual pitch should be equal, resulting in a diagonal line. In all experiments we considered, a high density of points around the diagonal line can be observed for mid-range pitches (i.e., close to the average pitch of the speakers). This indicates that the expected and actual pitch values are similar in this range, and therefore, the controllability of pitch works well. Comparing the distribution of pitch values of the experiments with unnormalized "orig" features (Fig. 2a) and globally normalized "normGlob" features (Fig. 2b), it can be seen that the latter exhibits inferior accuracies for low frequencies in female voices, explaining the high error for female voices in Tab. 2 (experiment b). The model trained only on data from female speakers (experiment f, see Fig. 2c left) is not able to reconstruct low frequencies accurately. Similarly, the model trained only on data from male speakers (experiment g, see Fig. 2c right) has problems with high frequencies. In summary, the plots in Fig. 2 confirm the result from Tab. 2, indicating that models with unnormalized ("orig") pitch features and a diverse training dataset achieve comparably high pitch accuracies for a large range of pitch values.

4 Conclusions

We trained acoustic models with the capability to control the prosody explicitly and examined the effects of feature normalization on the pitch ranges in the synthesized speech. In contrast to the common assumption of previous works, our results clearly show that the original feature range yields the lowest error when using concatenation. Furthermore, we can confirm that the pitch shifting capabilities per speaker benefit from the presence of other speakers with diverse pitch ranges in the training data.

Acknowledgments

This work was supported by the Free State of Bavaria in the DSAI project. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

References

[1] REN, Y., 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 *Proceedings of the International Conference on Learning Representations (ICLR)*. virtual, Austria, 2021.

[2] ŁANCUCKI, A.: *FastPitch: Parallel text-to-speech with pitch prediction*. In *Proceed-*

ings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 6588–6592. Toronto, Canada, 2021.

[3] REN, Y., Y. RUAN, X. TAN, T. QIN, S. ZHAO, Z. ZHAO, and T. LIU: *FastSpeech: Fast, robust and controllable text to speech*. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3165–3174. Vancouver, Canada, 2019.

[4] SANG, D. V. and L. X. THU: *FastTacotron: A fast, robust and controllable method for speech synthesis*. In *Proceedings of the International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. Hanoi, Vietnam, 2021.

[5] SCHÄFER, C., O. MCCARTHY, and CONTRIBUTORS: *ForwardTacotron*. `https://github.com/as-ideas/ForwardTacotron`, 2020.

[6] MOHAN, D. S. R., V. HU, T. H. TEH, A. TORRESQUINTERO, C. G. R. WALLIS, M. STAIB, L. FOGLIANTI, J. GAO, and S. KING: *Ctrl-P: Temporal Control of Prosodic Variation for Speech Synthesis*. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 3875–3879. Brno, Czech Republic, 2021.

[7] SHEN, J., R. PANG, R. J. WEISS, M. SCHUSTER, N. JAITLY, Z. YANG, Z. CHEN, Y. ZHANG, Y. WANG, R. RYAN, R. A. SAUROUS, Y. AGIOMYRGIANNAKIS, and Y. WU: *Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions*. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4779–4783. Calgary, Canada, 2018.

[8] BAKHTURINA, E., V. LAVRUKHIN, B. GINSBURG, and Y. ZHANG: *Hi-fi multi-speaker english tts dataset*. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*. Brno, Czech Republic, 2021.

[9] BONAFONTE, A., H. HÖGE, I. KISS, A. MORENO, U. ZIEGENHAIN, H. VAN DEN HEUVEL, H. HAIN, X. S. WANG, and M. GARCIA: *TC-STAR: Specifications of language resources and evaluation for speech synthesis*. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 311–314. Genoa, Italy, 2006.

[10] ZALKOW, F., P. GOVALKAR, M. MÜLLER, E. A. P. HABETS, and C. DITTMAR: *Evaluating speech–phoneme alignment and its impact on neural text-to-speech synthesis*. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1–5. Rhodes Island, Greece, 2023.

[11] KIM, J. W., J. SALAMON, P. LI, and J. P. BELLO: *CREPE: A convolutional representation for pitch estimation*. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 161–165. Calgary, AB, Canada, 2018.

[12] ZALKOW, F., P. SANI, M. FAST, J. BAUER, M. JOSHAGHANI, K. KAYYAR, E. A. P. HABETS, and C. DITTMAR: *The AudioLabs system for the Blizzard Challenge 2023*. In *Proceedings of the Blizzard Challenge Workshop*, pp. 63–68. Grenoble, France, 2023.

[13] MUSTAFA, A., N. PIA, and G. FUCHS: *StyleMelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization*. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6034–6038. Toronto, Canada, 2021.