# Low-Resource Text-to-Speech Using Specific Data and Noise Augmentation

Kishor Kayyar Lakshminarayana*, Christian Dittmar*, Nicola Pia*, Emanuël Habets†
*Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany
†International Audio Laboratories Erlangen, Germany

***Abstract*—Many neural text-to-speech architectures can synthesize nearly natural speech from text inputs. These architectures must be trained with tens of hours of annotated and high-quality speech data. Compiling such large databases for every new voice requires a lot of time and effort. In this paper, we describe a method to extend the popular Tacotron-2 architecture and its training with data augmentation to enable single-speaker synthesis using a limited amount of specific training data. In contrast to elaborate augmentation methods proposed in the literature, we use simple stationary noises for data augmentation. Our extension is easy to implement and adds almost no computational overhead during training and inference. Using only two hours of training data, our approach was rated by human listeners to be on par with the baseline Tacotron-2 trained with 23.5 hours of LJSpeech data. In addition, we tested our model with a semantically unpredictable sentences test, which showed that both models exhibit similar intelligibility levels.**

***Index Terms*—speech-synthesis, tacotron, text-to-speech, low-resource**

## I. Introduction

Modern neural text-to-speech (TTS) architectures such as Tacotron-1 and 2 [1], [2] require large quantities of paired text and high-quality speech recordings for each speaker [3] to synthesize near-natural speech from text input. For example, the popular LJ Speech [4] database, used extensively by the TTS community, has 23.5 hours of single-speaker samples. Collection and annotation of such a large database is resource-intensive, cumbersome, and expensive. Hence, there is a need for a TTS model that can be trained with a limited amount of paired text and speech data.

For well-resourced languages like English, there are a few publicly available TTS databases with samples from multiple speakers. Hence, transfer learning with these databases often complements training with a low-resource speaker like in [1], [5]. It is shown in [3] that transfer learning also works across languages. However using transfer learning could result in an unintentional transfer of speaking style or accent to the target speaker [6].

Data augmentation has been successfully used in neural network applications such as automatic speech recognition [7] and speaker verification [8]. Recently, multiple data augmentation techniques have been proposed for TTS as well. For example, [9], [10] use the CopyCat [11] voice conversion (VC) model to create augmentation samples. Articles like [12]–[14] use a teacher TTS to generate augmentation samples. This means that both the VC and TTS augmentation methods require a different neural architecture to be trained to produce the augmentation samples. Alternatively, data augmentation by changing the pitch of recorded speech was explored in [15], [16]. These were multi-speaker TTS scenarios using large training data sets of the order of five hours per speaker. Alternatively, [17] used unpaired speech and text data for augmentation, the collection of which is also time consuming. A recent article [18] explores data augmentation using re-ordering of the text-speech pairs, which requires meticulous preprocessing.

This paper proposes a method for single-speaker low-resource TTS training using specific data and noise augmentation. To prevent the noise augmentations from degrading the output quality, we extend the Tacotron-2 [2] architecture with additional augmentation embeddings. Although Tacotron is not the latest TTS architecture, it remains a competitive baseline in many recent studies, e.g., [14], [18]. In contrast to existing transfer learning approaches, the proposed approach does not require a pre-trained model. We also provide the specifics of the training data, which can be applied to any new voice or language to reduce the time and effort to collect the data.

Listening tests show that our approach achieves statistically equivalent Mean Opinion Scores (MOS) in comparison to a baseline Tacotron-2 model trained with the complete 23.5 hours of LJ-Speech [4] data, while ours only uses a 2-hours subset of the same corpus. We further demonstrate that the two approaches generate similar text and speech alignment, which is a critical requirement for synthesis. Additionally, we verified our proposed approach with objective intelligibility metrics.

## II. Proposed Method

We use the popular auto-regressive Tacotron-2 [2] to implement the low-resource TTS. The architecture with the proposed augmentation embedding extension is shown in Fig.1. Tacotron-2 converts an input text sequence to an output mel-spectrogram (mel) sequence via a sequence-to-sequence architecture.

Tacotron-2 uses an attention model to learn the alignment between each input text token and the corresponding mel frames. Typically, this attention model gives higher weights to the input at the current time. Hence, these attention weights indicate the alignment of the input text and output mel frames. Critically, Tacotron-2 cannot synthesize intelligible speech if it fails to learn proper alignment. Smaller training datasets, in a low-resource case, often result in memorization. i.e., only

[†] A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

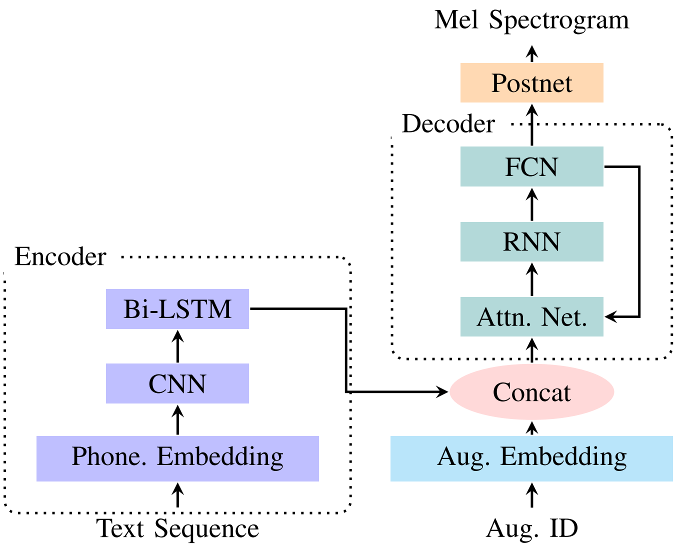


Fig. 1: Block diagram of Tacotron-2 with proposed noise augmentation embedding. Here, Bi-LSTM is Bidirectional Long Short Term Memory, CNN is convolutional neural network (NN), RNN is recurrent NN and, FCN is Fully Connected Network.

the text sequences used for training can be synthesized, while unseen textual inputs lead to garbled speech output.

The following sections describe our proposed methods and extensions to enable better alignment learning with limited training data.

### *A. Training Data Specification*

Like all neural networks, the training process of the Tacotron-2 model is carried out in batches. This means that the update of the weights happens once per batch by averaging the loss and deriving the corresponding weight update. In a typical training session, the samples used in a batch are randomly chosen, resulting in a wide range of sample durations. Additionally, this batch processing requires zero padding, which is expected to be discarded by the attention model. The attention model gets a single update over these varying durations. For a low-resource database, this effect and the small number of available batches result in poor alignment between text and mels. Though reduced batch size may reduce the duration variation, it makes the gradient noisier and causes memorization.

In this paper, we propose to reduce the variation within the training batches by providing a specification for the low-resource training data of the order of two hours. We recommend keeping the speech sample durations in the training dataset as close to one another as possible. For this purpose, we could use sentences or sentence segments (clauses) of similar length as the training data. A preprocessing step of splitting long sentences into short segments and otherwise using short sentences would achieve this. Further, this duration specification criterion should be used with the default consideration of having a broad, balanced phoneme distribution across the training set.

These similar-length training samples help the attention network learn better alignment between the text and speech. Moreover, the need for zero padding becomes minimal, offering an easier learning task for the attention network. Further, the ability of the model to synthesize longer sentences is not compromised since the component phoneme durations are not affected by the sentence being long. As a tradeoff, the prosodic interdependencies across long sentences are not learned using the current approach due to the use of short segments.

### *B. Noise Augmentation*

We propose using noise augmentation to increase the number of samples available for training in a low-resource setting. Such an augmentation can create multiple samples with the same phoneme/speech content but significantly different mels. Using such noise-augmented samples can lead to degradation of quality, which is then avoided using the augmentation labels.

A speaker label is often provided along with the text as an input to TTS models to synthesize different voices, e.g., [1], [19]. These labels help to generate neural embeddings as speaker representations. Similarly, we propose neural augmentation embeddings using specified augmentation identifiers (Aug. ID.). These are concatenated to the output of the Tacotron-2 encoder as shown in Fig. 1. These neural augmentation embeddings are the common representation (multi-dimensional vector) learned across the samples of the distinct augmentation sets. For every augmentation set, a stationary noise with a known statistical distribution at a specific Signal-to-Noise Ratio (SNR) is added to each training sample. We used three kinds of stationary noise types, empirically defined, each at different SNRs. The three noise types were assigned to three Aug. IDs. The original data is associated with the ”clean” Aug. ID, later used for inference. We found that using three different noise types for augmentation resulted in the same outcome, regardless of the specific noise types used. We also found that adding more than three noise augmentations would increase the training time without significant performance gain.

Intuitively, the augmentation embedding layer learns the common properties of the noise (or no) augmentation. Hence the rest of the model parameters are influenced by the relation between the input phonemes and the corresponding mels.

## III. Experimental Setup

### *A. Datasets*

We simulated the training data specification proposed in Section II-A using subsets totaling 2 hours from the LJ Speech database. Through experimentation, we discovered that using less than 2 hours of data does not produce usable speech, even with noise augmentation. As a baseline, we randomly selected 2 hours of data from the prompts resulting in a 2H_RS (random set) dataset. Then, we arranged the whole 23.5-hour dataset in the order of durations and picked sentences (starting with the shortest duration) totaling 2 hours, resulting in 2H_IS (informed set) dataset. The arrangement ensured that the samples were of short duration and that the durations in a training batch were close. We then used both 2H_RS and 2H_IS datasets with and without noise augmentation.

We used white Gaussian noise (WGN) at 25 dB SNR, United States of America Standards Institute (USASI) standard noise at 15 dB SNR, and noise simulated from the noise power spectral density of a Knowles EM-3346 electret microphone (sensor noise) at 20 dB SNR as the three additive noises [20]. We used the Voicebox toolbox [20] for the noise addition. This determines the SNR based on the active speech power following the ITU-T P.56 recommendation. We found that specific noise types and SNR levels did not affect the results as long as there were three different augmentations.

### B. Neural Vocoder

We used a pre-trained StyleMelGAN [21] as the neural vocoder for converting the predicted mels to speech. The vocoder was trained for German synthesis with PAVOQUE [22], CSS_10 [23], BITS [24], and a proprietary two-speaker (one male and one female) dataset. This vocoder achieved good quality in listening tests conducted for English speech with copy synthesis and hence was used for the current evaluation. As an important side note, the vocoder can be trained with unlabelled speech data which is easier to collect and more readily available.

### C. Experiments

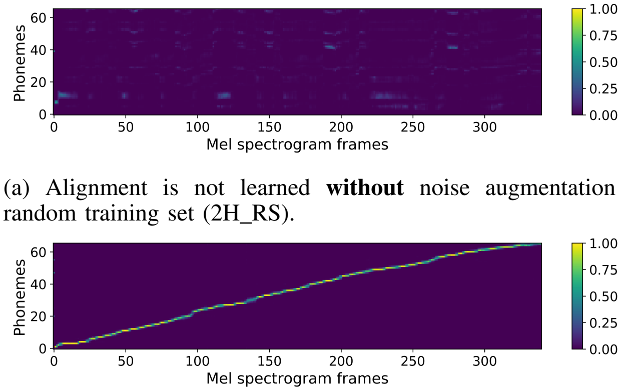
We simulated a low-resource scenario using LJ Speech data. We used the Tacotron-2 model trained with the full 23.5 hour LJ Speech dataset as the baseline. We then used the two separate low-resource datasets mentioned in Sec. III-A for training. These are the 2H_RS, a randomly selected low-resource dataset and the 2H_IS, which simulates the proposed training data specification. These datasets were used to train the baseline Tacotron-2 model resulting in 2H_RS and 2H_IS models, respectively. We further trained the proposed model shown in Fig.1 using these datasets and their noise-augmented versions, separately for RS and IS sets, resulting in 2H_RS_NA and 2H_IS_NA models. The training was done for 310K iterations, each with a batch size 24. Audio samples are shared at https://s.fhg.de/lrtts.

Since the results from a single dataset training could be attributed to chance, we further investigated low-resource speakers 'bdl' (male) and 'slt' (female) from the CMU ARCTIC English speech synthesis database [25]. The results here were comparable. i.e., we could synthesize intelligible and good quality speech from 'bdl'/'stl' samples only after using specific samples by excluding "long" samples and using augmentation. All these experiments were different single-speaker training sessions.

## IV. Results

### A. Text to Speech Alignment Learning

We analyzed the learned attention weights across the input phoneme tokens and output mel spectrogram frame durations for each training configuration to verify our experiments' viability. In Fig. 2, weights are plotted for a representative speech, LJ002-0114.wav, from the LJ Speech dataset, for two configurations. We see a diagonal tendency in Fig. 2b, which indicates that the alignment between phonemes and mel frames has been learned with the 2H_IS_NA model. In contrast, Fig. 2a shows that the attention weights are spread out flat, and no reasonable alignment is learned with the 2H_RS model.



(a) Alignment is not learned **without** noise augmentation with random training set (2H_RS).

(b) Alignment is learned **with** noise augmentation and informed training set (2H_IS_NA).

Fig. 2: Learned attention weights for LJ002-0114.wav from the LJ Speech dataset in two training sessions.

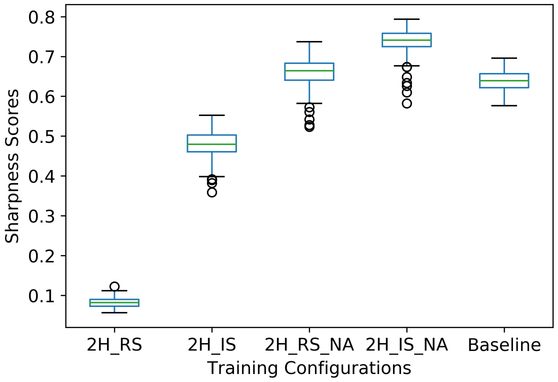


Fig. 3: Sharpness scores of the learned attention weights across 174 prompts for the high-resource baseline and different low resource configurations.

Suppose the attention model learns a good alignment between the phonemes and the corresponding mel frames. In that case the learned attention weight should be maximum (close to 1.0) at the aligned position and zero elsewhere. We define a sharpness score (inspired by [26]) as the mean of the maximum attention weights across each mel frame. A sharpness score close to 1.0 indicates that the learned alignment is good. The boxplot of the sharpness scores of the learned attention weights for 174 sentences for different training configurations is plotted in Fig. 3. This figure shows that the random training data selection (RS) does not learn proper alignment, and the specific short duration set (IS) improves the alignment quite a bit. Both the noise augmented (NA) versions have learned good alignment, and the specific set with noise augmented training gets even sharper alignments than the baseline high resource training with the entire 23.5 hour database.

TABLE I: Evaluation metrics across the different simulated low resource configurations. MOS ratings are listed with 95% confidence intervals. The ASR based SUS WER are given in percent.

| LJ Speech Set | MOS (↑) | SUS-WER (↓) |
|---|---|---|
| 2H_RS | $1.21 \pm 0.04$ | 133.1 |
| 2H_RS_NA | $2.17 \pm 0.09$ | 76.5 |
| 2H_IS | $3.02 \pm 0.09$ | 38.2 |
| 2H_IS_NA | $\mathbf{3.98 \pm 0.09}$ | 19.6 |
| Baseline (23.5H) | $3.97 \pm 0.09$ | **18.5** |

### B. Quality Evaluation

We used subjective Mean Opinion Score (MOS) tests and objective intelligibility tests based on Semantically Unpredictable Sentences (SUS) for evaluation. These tests are described below.

The MOS tests were conducted using Absolute Category Rating as per P.808 [27] using WebMUSHRA [28] with 14 expert listeners from our laboratory having no reported hearing impairments, with an average age of 32.5 years. We used the first four lists from Harvard sentences [29] with 40 sentences. At least 12 listeners rated each sentence in all the conditions.

The SUS test was proposed as an objective measure of intelligibility for TTS models in [30]. We synthesized all the 100 SUS texts which were part of the Blizzard challenge of 2005 [31] with the TTS model under test. These were then transcribed with a pre-trained Speechbrain automatic speech recognition (ASR) [32] recipe. The word error rate (WER) was measured using the Python JiWER package. A lower WER indicates better intelligibility.

Table I gives the results from both tests, showing that the proposed model with 2 hours of specific training data and noise augmentation (2H_IS_NA) performs almost as well as the baseline version trained with 23.5 hours of data. The other configurations are progressively worse. Further, we can also see that the model trained using a random low resource set without noise augmentation (2H_RS) cannot synthesize any intelligible speech as the WERs here are higher than 100 percent.

We also evaluated the models with long sentences, e.g., the 46-word sentence from 'There There' by Tommy Orange [33]. The output for this sentence with the proposed low-resource approach and the baseline approach was similar in terms of intelligibility during informal listening. Both had WER of 9% with the ASR test.

## V. Conclusion

We implemented a low-resource single-speaker TTS model with minor modifications to the Tacotron-2 architecture. We proposed a training approach that does not require a separate pre-trained model and does not suffer from the accent and style transfer issues commonly present in multi-speaker, multi-language approaches. The proposed approach uses only 2 hours of specific data and noise augmentation for training. Further, augmentation identifiers are used to learn augmentation embeddings. Using specific training data and noise augmentation improved the learned text and speech alignment and thereby the synthesis. We demonstrated that the quality of speech synthesized by the proposed approach trained with only 2 hours of specific data is comparable to the speech synthesized by the baseline architecture trained with 23.5 hours of data using subjective tests. Further, objective intelligibility tests were conducted that support this finding. These insights can be used to train TTS models using much less data and still achieve natural quality speech synthesis. They might be used to train TTS models for dialects and languages for which little data is available. In the future, we plan to improve the specification process to account for phoneme distribution balance and study the effects of the current approach on the prosody of sentences.

## VI. Acknowledgements

Parts of this work have been supported by the SPEAKER project (FKZ 01MK20011A), funded by the German Federal Ministry for Economic Affairs and Climate Action. In addition, this work was supported by the Free State of Bavaria in the DSAI project.

## References

[1] A. Gibiansky, S. Ö Arık, et al., "Deep Voice 2: Multi-speaker neural text-to-speech," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[2] J. Shen, R. Pang, et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[3] J. Xu, X. Tan, et al., "LRSpeech: Extremely low-resource speech synthesis and recognition," in *Proc. ACM Intl. Conf. on Knowledge Discovery & Data Mining (SIGKDD)*, 2020, pp. 2802–2812.

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

[5] N. Tits, K. El Haddad, and T. Dutoit, "Exploring transfer learning for low resource emotional TTS," in *Proc. SAI Intelligent Systems Conference*. Springer, 2019, pp. 52–60.

[6] J. Latorre, C. Bailleul, et al., "Combining speakers of multiple languages to improve quality of neural voices," in *Proc. ISCA Speech Synthesis Workshp*, 2021, pp. 37–42.

[7] D. S. Park, W. Chan, et al., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[8] Z. Wu, S. Wang, et al., "Data Augmentation Using Variational Autoencoder for Embedding Based Speaker Verification," in *Proc. Interspeech*, 2019, pp. 1163–1167.

[9] R. Shah, K. Pokora, et al., "Non-autoregressive TTS with explicit duration modelling for low-resource highly expressive speech," in *Proc. ISCA Speech Synthesis Workshp*, 2021, pp. 96–101.

[10] G. Huybrechts, T. Merritt, et al., "Low-resource expressive text-to-speech using data augmentation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6593–6597.

[11] S. Karlapati, A. Moinet, et al., "CopyCat: Many-to-many fine-grained prosody transfer for neural text-to-speech," in *Proc. Interspeech*, 2020, pp. 4387–4391.

[12] M. Sharma, T. Kenter, and R. Clark, "StrawNet: Self-Training WaveNet for TTS in Low-Data Regimes," in *Proc. Interspeech*, 2020, pp. 3550–3554.

[13] MJ. Hwang, R. Yamamoto, et al., "TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6598–6602.

[14] E. Song, R. Yamamoto, et al., "TTS-by-TTS 2: Data-Selective Augmentation for Neural Speech Synthesis Using Ranking Support Vector Machine with Variational Autoencoder," in *Proc. Interspeech*, 2022, pp. 1941–1945.

[15] E. Cooper, CI. Lai, et al., "Can Speaker Augmentation Improve Multi-Speaker End-to-End TTS?," in *Proc. Interspeech*, 2020, pp. 3979–3983.

[16] B. Lőrincz, A. Stan, and M. Giurgiu, "Speaker verification-derived loss and data augmentation for dnn-based multispeaker speech synthesis," in *Proc. IEEE-SPS European Signal Processing Conf.*, 2021, pp. 26–30.

[17] Y. Chung, Y. Wang, et al., "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6940–6944.

[18] M. Lajszczak, A. Prasad, et al., "Distribution augmentation for low-resource expressive text-to-speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8307–8311.

[19] Y. Jia, Y. Zhang, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," 2018, vol. 31.

[20] M. Brookes, "Voicebox: Speech processing toolbox for matlab," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 2000.

[21] A. Mustafa, N. Pia, and G. Fuchs, "StyleMelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6034–6038.

[22] I. Steiner, M. Schröder, and A. Klepp, "The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech," *Proc. Phonetik & Phonologie*, vol. 9, 2013.

[23] K. Park and T. Mulc, "CSS10: A collection of single speaker speech datasets for 10 languages," *Proc. Interspeech*, pp. 1566–1570, 2019.

[24] T. Ellbogen, F. Schiel, and A. Steffen, "The BITS speech synthesis corpus for German," in *Proc. (ELRA) Intl. Conf. on Language Resources and Evaluation*, 2004.

[25] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. ISCA Speech Synthesis Workshp*, 2004.

[26] C. Schäfer, "ForwardTacotron," https://github.com/as-ideas/ForwardTacotron, 2020.

[27] ITUT Rec, "P. 808, subjective evaluation of speech quality with a crowdsourcing approach," *ITU-T, Geneva*, 2018.

[28] M. Schoeffler, S. Bartoschek, et al., "webMUSHRA—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[29] E.H. Rothauser, W.D. Chapman, et al., "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, 1969.

[30] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," vol. 18, no. 4, pp. 381–392, 1996.

[31] A Black and K. Tokuda, "The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common databases," in *Proc. Interspeech*, 2005, pp. 77–80.

[32] M. Ravanelli, T. Parcollet, et al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[33] Henneke, "How to captivate readers with a dazzling loooong sentence," https://www.enchantingmarketing.com/how-to-write-a-long-sentence/, 2021.