13th edition of the Speech Synthesis Workshop
24-27 August 2025, Leeuwarden, The Netherlands

Explicit Emphasis Control in Text-to-Speech Synthesis

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

[1]Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany
[2]International Audio Laboratories Erlangen, Germany

judith.bauer@iis.fraunhofer.de, frank.zalkow@iis.fraunhofer.de,
meinard.mueller@audiolabs-erlangen.de, christian.dittmar@iis.fraunhofer.de

Abstract

Recent text-to-speech (TTS) systems are able to generate synthetic speech with high naturalness. However, the synthesized speech usually lacks variation in emphasis. Since it is well-known that emphasizing different words can alter a sentence's meaning, it is desirable to extend TTS models to include the ability for emphasis control, i.e., the option to indicate during synthesis which words should carry special emphasis. In this work, we realize such functionality by automatically annotating TTS training datasets with emphasis scores and modifying the TTS model to use these scores during training. In particular, we propose a new architecture for emphasis detection and compare its suitability for TTS with existing emphasis detectors. We introduce an extension for the ForwardTacotron TTS model and train multiple versions of the model with scores from the different emphasis detectors. Finally, we compare the naturalness and the perceived emphasis of speech synthesized by the models.

Index Terms: Text-To-Speech Synthesis, controllable, prosody, emphasis, prominence

1. Introduction

Modern text-to-speech (TTS) systems are able to synthesize highly natural speech from text. However, many systems lack the option of fine-grained control over the generated speech. Depending on the application scenario, it is desirable to have control over certain aspects of the generated speech signals, e.g., controllable prosody. In spoken language, the meaning of a sentence does not only depend on its textual content but also on the emphasis (also referred to as focus or prominence) [1]. Emphasis can be used to highlight new information, as opposed to information that is already known from the context. It is also applied when correcting wrong information with the contrasting true information, as in the following example:

> Question 1: "Did she buy four apples?"
>
> Question 2: "Did she buy five bananas?"
>
> Answer: "She actually bought five apples."

If the answer responds to the first question, the word "five" would be emphasized, and if it answers the second question, the emphasis would be on the word "apples." Since emphasis is an important aspect of spoken language, controllable emphasis, i.e., the option to deliberately modify the emphasis of a word during synthesis, is a desirable feature in TTS.

In this work, we propose a new approach for training TTS models with controllable emphasis by leveraging existing datasets with acted emphasis (i.e., recordings with strong emphasis on a previously defined word). Using those datasets, we train an emphasis detector model, which learns to predict emphasis scores from text and audio features. Throughout this paper, the term "emphasis score" refers to a real-valued scalar assigned to a word, describing how strongly emphasized the corresponding word was pronounced.

We apply this model and other pretrained emphasis detectors to generate word-wise emphasis scores for the datasets used to train the TTS model. Our TTS model contains prosody predictors for fundamental frequency (referred to as pitch in the following), energy (loudness), phoneme durations (speech rate), and voicing confidence (saliency of the pitch). Since these features are related to emphasis, the prosody predictions should vary depending on the desired emphasis. Therefore, the emphasis scores are given as input to the prosody predictors, allowing the emphasis information to steer the prosody estimates. During inference, no precomputed emphasis scores are available, hence we use values determined from dataset statistics as explained later. By modifying these word-wise values, the expressed emphasis can be gradually adjusted, realizing controllable emphasis. We provide audio examples on an accompanying website[1].

Our proposed system is based on emphasis scores predicted by an emphasis detector. In previous work, several approaches for detecting emphasis in speech recordings exist. As part of their work on emphasis transfer in speech, de Seyssel et al. [2] propose to finetune a cross-lingual speech representation model for emphasis classification. Morrison et al. [3] suggest a convolutional emphasis detection model consisting of an encoder part operating on the frame level and a word-level decoder. In our experiments, we use these two approaches in addition to our proposed emphasis detector. Specifically, we evaluate their influence on training TTS models with controllable emphasis and compare these TTS models to a TTS model relying on rule-based prosody modifications [4].

An alternative to our approach of using emphasis values based on dataset statistics during inference is adding an emphasis predictor to the TTS model. This emphasis predictor can be trained together with the TTS model and learns to estimate emphasis scores from textual information. During inference, the predicted scores can be modified to change the synthesized emphasis. While the computation of emphasis scores at inference time differs between these prediction-based approaches and our statistics-based approach, both allow for manually changing the default scores at inference time. Seshadri et al. [5] propose an emphasis predictor trained with variance-based and wavelet-based features, which steer predictors for pitch, energy, and duration. A similar approach is taken by Niu and Silamu [6], who use an emphasis predictor trained with wavelet-based features. While these emphasis predictors receive (encoded) phoneme

[1]https://s.fhg.de/explicit-emphasis-control

10.21437/SSW.2025-4

information, Zhong et al. [7] train an emphasis predictor that receives additional linguistic information as input. Malisz et al. [8] propose an approach where emphasis features are predicted from high-dimensional linguistic features and used as input to a small feed-forward acoustic model.

Another approach to employ emphasis information in a TTS model was proposed by Liu et al. [9], who add an emphasis embedding to a forward attention module. Suni et al. [10] propose to add the emphasis information by augmenting the textual input. Their procedure is partly similar to ours, as they automatically generate emphasis scores and use them for training a TTS model without an emphasis prediction module. There are two main differences to our work: (a) Their scores are computed with a method based on the continuous wavelet transform, instead of being predicted by a neural emphasis detection model, and (b) their emphasis scores are subsequently discretized resulting in three emphasis classes, while we use real-valued scores predicted by detector models, allowing for continuous emphasis control during inference.

Focusing more directly on prosody features, other authors integrate emphasis by including prosody predictors and modifying their estimates. Raitio et al. [11] propose a model with hierarchical prosody predictors and suggest to modify the pitch range and phoneme durations for emphasized words. Also using prosody at different hierarchies, Shechtman et al. [12] estimate pitch spread and average phoneme durations on word and sentence level. These estimates can be modified at inference time to generate emphasis. Joly et al. [4] use a simpler approach, which only requires modification of duration predictions.

With the recent rise of large language models, integrating those with speech synthesis holds promise for automatically generating speech with suitable emphasis. However, we propose a more lightweight approach for explicit emphasis control, requiring smaller datasets and models, and less computational resources. Furthermore, our approach offers two additional possibilities: to realize emphasis in single-sentence prompts without a given context and to adjust the degree of emphasis according to users' requirements.

Our main contributions are:

- We propose a new architecture for an emphasis detector.
- Using our emphasis detector and other methods for predicting emphasis scores, we train multiple variants of a TTS model with controllable emphasis.
- To the best of our knowledge, this study is the first to compare TTS models trained with labels from multiple emphasis detectors. We evaluate the models in terms of perceived naturalness and emphasis strength. Additionally, we investigate the effect of emphasis levels on the prosody predictions.
- We show that the TTS system trained with emphasis scores from our new detector model performs better at synthesizing words with strong emphasis compared to systems trained with scores from other detectors.

2. Method

2.1. Datasets and Extracted Features

In the following experiments, we use the datasets shown in Table 1. DB-TTS contains utterance-aligned text transcriptions and audio recordings with mostly neutral speaking style from five English speakers. No ground-truth emphasis scores are available for this dataset. The datasets DB-E (subset of Expresso dataset [13]) and DB-S (subset of SIWIS French Speech Synthesis Database [14]) contain recordings with acted emphasis, the corresponding text transcriptions, and ground-truth binary emphasis scores.

All audio recordings are resampled to a sampling frequency of 22 050 Hz. From the recordings, we extract mel spectrograms with 80 bands, a hop size of 256 samples, a block size of 1024 samples, and an upper frequency limit of 8000 Hz. The text transcriptions are converted to phoneme transcriptions through look-up in an English pronunciation dictionary, supplemented by espeak[2] for out-of-vocabulary words. For each utterance, we compute a temporal alignment between the mel spectrogram and the corresponding phoneme transcription using a CTC-based aligner model [15]. This alignment specifies the number of mel frames belonging to each phoneme and can be utilized to aggregate prosody features. To this end, we extract pitch and voicing confidence by analyzing the speech recordings with CREPE [16], and aggregate these features to phoneme-wise values using the alignment information. Energy information is derived by computing the L2-norm of the mel spectrogram frames and is aggregated in a similar fashion. We refer to the prosody-related features for pitch, confidence, energy, and phoneme duration on the phoneme level as p_{pitch}, p_{conf}, p_{energy}, and p_{dur}, respectively.

2.2. Emphasis Detectors

The goal of emphasis detection is to estimate one emphasis score for each word of an utterance, using features such as audio features or text information. Datasets with emphasis annotations often encode emphasis with binary labels, where 0 corresponds to no emphasis and 1 indicates that a word is emphasized. Since the nature of emphasis in spoken language is not actually binary, emphasis detection models trained on such datasets do not reproduce the binary labels exactly, resulting in continuous scores in the range $[0, 1]$. These real-valued emphasis scores are useful for training TTS models, because exposing the model to continuous scores during training results in models capable of gradual emphasis controllability at inference time. We experiment with different approaches for emphasis detection.

RNN-based detector: We propose an emphasis detection model based on a recurrent neural network (RNN). The model receives textual features and prosody features p_{pitch}, p_{conf}, p_{energy}, and p_{dur} as input. The model architecture consists of linear layers and bidirectional LSTM layers. After a final sigmoid activation function, we receive one real-valued emphasis score per word. For a detailed description of the model, see Sec. 3.1.1.

XLS-R-based detector: As part of their work on a benchmark for evaluating emphasis in speech-to-speech models, de Seyssel et al. [2] proposed a model[4,5] for emphasis detection from speech waveforms. Therefore, the authors finetune a cross-lingual speech representation (XLS-R) model to perform frame-wise emphasis classification. Using the frame-wise classification, a word is considered emphasized if more than 50 percent of the frames belonging to the word are predicted as

[2] https://espeak.sourceforge.net/

[3] https://keithito.com/LJ-Speech-Dataset

[4] https://github.com/facebookresearch/emphassess

[5] Checkpoint: https://dl.fbaipublicfiles.com/speech_expressivity_evaluation/EmphAssess/EmphaClass/EmphaClass-en.tar.gz

Table 1: *Overview of datasets used for training the emphasis detector and the acoustic model*

Dataset	Language	Speakers	Duration (in h)	Source	Annotated emphasis
DB-TTS	en	"female1"	23.03	Hi-Fi TTS (speaker 92) [17]	✗
		"female2"	2.22	internal	✗
		"female3"	23.42	LJSpeech[3]	✗
		"male1"	8.60	TC-STAR [18]	✗
		"male2"	2.14	internal	✗
DB-E	en	2 female & 2 male	0.41	Expresso [13] (subset)	✓
DB-S	fr	1 female	0.48	SIWIS French [14] (subset)	✓

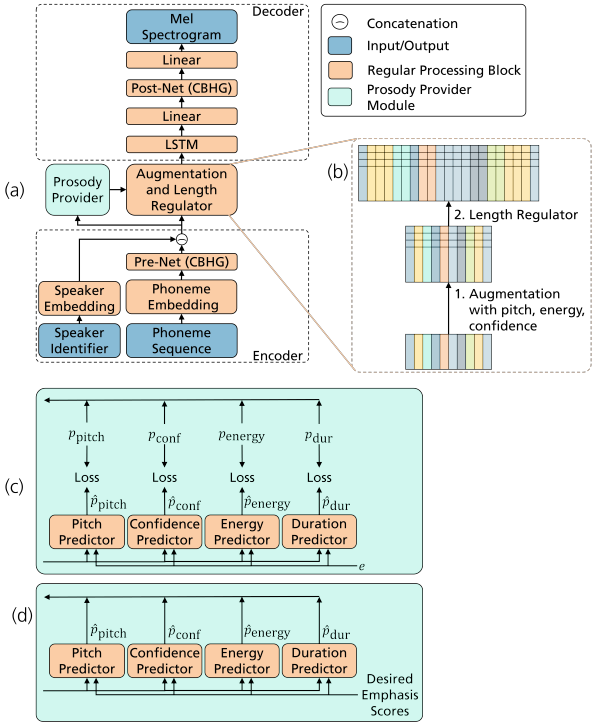

Figure 1: *Overview of the acoustic model used in this paper. Subfigure (a) shows the architecture of the acoustic model. Subfigure (b) shows details of the Augmentation and Length Regulator step. Subfigures (c) and (d) illustrate the details of the prosody provider during training and inference, respectively.*

emphasized. Since we require real-valued scores for our experiments, we use the percentage of frames that are classified as emphasized as an emphasis score.

Combination of RNN-based detector and XLS-R-based detector: We noticed that the predictions by the XLS-R-based detector have a low recall and a high precision compared to our RNN-based detector, which means that the XLS-R-based detector often assigned low scores to emphasized words. However, our RNN-based detector sometimes overpredicts the scores for non-emphasized words. To compensate for both models' weaknesses, we propose combining the two: If both methods assign low scores to a word (i.e., the XLS-R-based detector assigns zero and the RNN-based detector assigns < 0.5), we consider a word as non-emphasized and set the score to zero. Otherwise, the prediction by the RNN-based detector is used.

CNN-based detector: Morrison et al. [3] trained an emphasis predictor on crowdsourced emphasis annotations and published[6] a convolutional neural network (CNN), which uses convolution layers for a framewise encoding of mel spectrograms. Using a word-level time alignment, these encodings are downsampled to word-wise encodings by aggregating all frames associated with each word. From these encodings, a convolutional decoder predicts an emphasis score for each word.

2.3. TTS System with Emphasis Information

For the following experiments, we use a TTS system consisting of an acoustic model for predicting mel spectrograms and a vocoder model, which uses the mel spectrograms to synthesize speech. Our acoustic model is based on ForwardTacotron[7] with additional prosody predictor modules for pitch, energy, voicing confidence, and phoneme duration prediction, similar to the model by Zalkow et al. [19]. Fig. 1a gives an overview of the architecture, consisting of three parts: (1) encoder, (2) augmentation with prosody features and length regulator, and (3) decoder.

(1) Encoder: As shown in Fig. 1a, the model receives phoneme and speaker information as input, and the encoder generates an intermediate representation, keeping the feature sequence length the same as the phoneme sequence length.

(2) Augmentation with Prosody-Related Features and Length Regulator: During training (see Fig. 1c), a prosody provider module receives the encoder output and emphasis features e (per-phoneme emphasis scores obtained by replicating per-word scores). A set of predictors use these features to estimate pitch $\hat{p}_{\text{pitch}}$, confidence $\hat{p}_{\text{conf}}$, energy $\hat{p}_{\text{energy}}$, and duration $\hat{p}_{\text{dur}}$. To train the predictors, the mean-absolute-error loss between the predicted features and the ground-truth features p_{pitch}, p_{conf}, p_{energy}, and p_{dur} is computed. During training, the ground-truth prosody features are passed through by the prosody provider (teacher-forcing) and p_{pitch}, p_{conf}, and p_{energy} are concatenated to the encoder output. Using the duration information p_{dur}, the length regulator repeats features in a non-equidistant fashion along the temporal dimension to match the number of time frames in the target mel spectrogram (Fig. 1b). During inference (see Fig. 1d), no ground-truth prosody features are available, so the predicted features $\hat{p}_{\text{pitch}}$, $\hat{p}_{\text{conf}}$, $\hat{p}_{\text{energy}}$, and $\hat{p}_{\text{dur}}$ are used directly.

(3) Decoder: From the augmented and resampled intermediate representation, the decoder predicts a mel spectrogram.

[6] https://github.com/interactiveaudiolab/emphases

[7] https://github.com/as-ideas/ForwardTacotron

Table 2: *Median emphasis scores for words that are emphasized/non-emphasized in DB-E.*

Emphasis Detector	$\text{med}_{\text{nonemph}}$	med_{emph}
RNN	0.0447	0.9711
XLS-R	0.0000	0.9908
Combined RNN and XLS-R	0.0000	0.9711
CNN	0.1516	0.5513

The network is trained by computing the mean-absolute-error loss between the predicted and the ground-truth mel spectrogram.

For synthesizing speech from the mel spectrograms, we use the StyleMelGAN vocoder [20], which was pretrained on multiple speakers including the ones used for training the acoustic model.

3. Experiments

3.1. Experimental Details

3.1.1. Emphasis Detector Models

The RNN-based emphasis detection model mentioned in Sec. 2.2 is a neural network that receives prosody features p_{pitch}, p_{conf}, p_{energy}, and p_{dur} and text information in the form of articulatory features (similar to [21], i.e., one-hot encoding of different articulatory attributes, e.g., tongue position and vowel openness) as input. For each phoneme, each of these five features is projected to a 128-dimensional vector using a linear layer. These five vectors are then concatenated along the feature dimension, resulting in one 640-dimensional vector per phoneme, which is further processed using two linear layers with output dimensions of 512 and 256, respectively. To incorporate temporal relations between the phoneme-wise vectors, a bidirectional LSTM layer with an output dimension of 256 is applied to the sequence of vectors. After splitting the output features of the LSTM at the word boundaries, we apply another bidirectional LSTM (output dimension 256) for each word. The final hidden states are processed by a linear downprojection layer with a sigmoid activation function, resulting in a single score for each word. We train the model for 400 epochs with a batch size of 16 using the binary cross-entropy loss between the predicted emphasis scores and the binary ground-truth emphasis annotations. The model is trained with the Adam optimizer [22] and a linearly decreasing learning rate which starts at 10^{-5} and ends at $5 \cdot 10^{-6}$. As training datasets, we use DB-E and DB-S (see Table 1, duration <1h). In preliminary experiments, we found that it is important to balance the loss according to the ratio of non-emphasized words to emphasized words, since the number of emphasized words is much smaller than the number of non-emphasized words.

For the XLS-R [2] and the CNN-based [3] emphasis detection approaches, we use available pretrained models (trained on more data than our RNN-based model). Furthermore, the predictions of the RNN-based detector and the XLS-R-based detector are combined as explained in Sec. 2.2.

3.1.2. Acoustic Model Variants

The RNN, XLS-R, combined RNN with XLS-R, and CNN-based emphasis detectors are applied to DB-TTS (dataset for training the TTS models, see Table 1) to generate per-word emphasis scores for all utterances of the dataset. After replicating the obtained per-word emphasis scores to per-phoneme scores e, we train an acoustic model for each emphasis detection variant, each using their respective scores e to condition the prosody predictors as described in Sec. 2.3. We refer to the resulting acoustic models as $\mathcal{M}_{\text{RNN}}$, $\mathcal{M}_{\text{XLSR}}$, $\mathcal{M}_{\text{RNN+XLSR}}$, and $\mathcal{M}_{\text{CNN}}$. Furthermore, we train a model without emphasis scores $\mathcal{M}_{\text{NoEmph}}$. Joly et al. [4] suggest that emphasis can be modelled by scaling predicted phoneme durations with a constant factor (**D**uration **D**ilatation). Using $\mathcal{M}_{\text{NoEmph}}$, we realize this approach and refer to it as $\mathcal{M}_{\text{DD}}$. Finally, to obtain a lower anchor for our listening tests (described in Sec. 3.2), we use $\mathcal{M}_{\text{RNN}}$ with randomized emphasis scores outside of the range seen during training. This approach is denoted by $\mathcal{M}_{\text{Anchor}}$.

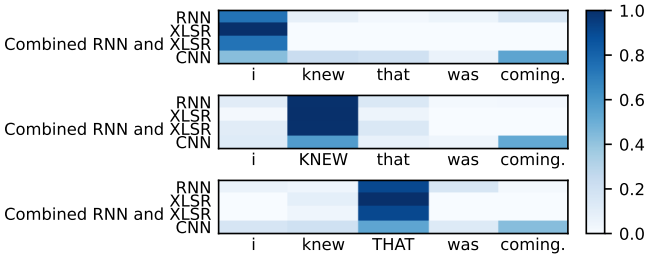

Figure 2: *Per-word emphasis predictions by emphasis detectors applied to three prompts from DB-E (dataset with annotated emphasis). Uppercase words indicate that the word was emphasized in the recording. The color intensity refers to the predicted emphasis scores.*

During training, the acoustic models use the replicated scores predicted by the respective emphasis detectors. Since our acoustic models have no means to predict emphasis scores during inference, we determine appropriate scores by applying the emphasis detection models to DB-E (see Fig. 2 for an example of scores predicted by the emphasis detectors). We propose to compute the median of the scores for non-emphasized words $\text{med}_{\text{nonemph}}$ and use this value as score for non-emphasized words during inference. For words that should be emphasized, we set the score to a higher value, e.g., by linear interpolation between $\text{med}_{\text{nonemph}}$ and the median of the scores for emphasized words med_{emph}

$$\text{med}_{\text{nonemph}} + \alpha \cdot (\text{med}_{\text{emph}} - \text{med}_{\text{nonemph}}), \qquad (1)$$

where a higher $\alpha \in \mathbb{R}_{\geq 0}$ corresponds to more emphasis. Table 2 shows the values of $\text{med}_{\text{nonemph}}$ and med_{emph} for the emphasis detection models. Due to the internal characteristics of $\mathcal{M}_{\text{XLSR}}$ and $\mathcal{M}_{\text{RNN+XLSR}}$, $\text{med}_{\text{nonemph}}$ is zero for these models, which makes their usage especially intuitive.

The acoustic model variants were trained with a batch size of 32 for 300k training iterations, using a learning rate of 10^{-4} until training iteration 150k and 10^{-5} afterwards. As optimizer, we use Adam [22]. We train the prosody predictors jointly with the acoustic model by combining the L1 loss values between predicted and ground-truth prosody values with the L1 loss values between predicted and ground-truth mel spectrograms. During training, we normalize the ground-truth pitch and energy values with the mean and standard deviation computed across all speakers in the dataset.

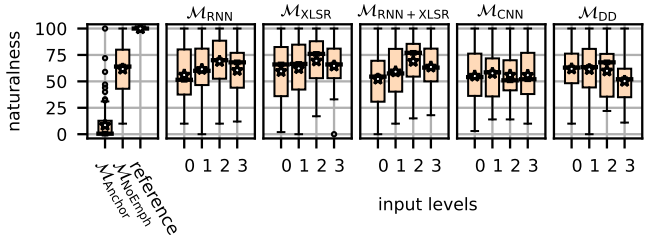

Figure 3: *Results of the subjective evaluation of naturalness. The left plot shows the scores for $\mathcal{M}_{\text{Anchor}}$, $\mathcal{M}_{\text{NoEmph}}$, and the reference. The other plots show the scores depending on the emphasis input levels for $\mathcal{M}_{\text{RNN}}$, $\mathcal{M}_{\text{XLSR}}$, $\mathcal{M}_{\text{RNN+XLSR}}$, $\mathcal{M}_{\text{CNN}}$, and $\mathcal{M}_{\text{DD}}$. Stars indicate the mean values.*

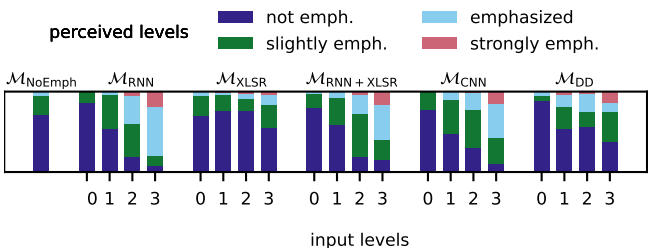

Figure 4: *Results of subjective evaluation of perceived emphasis depending on input emphasis level for $\mathcal{M}_{\text{NoEmph}}$, $\mathcal{M}_{\text{RNN}}$, $\mathcal{M}_{\text{XLSR}}$, $\mathcal{M}_{\text{RNN+XLSR}}$, $\mathcal{M}_{\text{CNN}}$, and $\mathcal{M}_{\text{DD}}$.*

Table 3: *Analysis of results from subjective evaluation of perceived emphasis. Pearson correlation coefficients and slopes of regression lines show the correlation between input emphasis levels and perceived emphasis levels.*

	$\mathcal{M}_{\text{RNN}}$	$\mathcal{M}_{\text{XLSR}}$	$\mathcal{M}_{\text{RNN+XLSR}}$	$\mathcal{M}_{\text{CNN}}$	$\mathcal{M}_{\text{DD}}$
corr.	0.7075	0.1624	0.5928	0.5764	0.3080
p-value	0.0000	0.0263	0.0000	0.0000	0.0000
slope	0.6168	0.1033	0.4789	0.4564	0.2486

3.2. Subjective and Objective Evaluation

When integrating controllable emphasis in TTS, it is desirable that the synthesized speech sounds natural and the intended emphasis is perceivable to the listener. Therefore, we conduct listening tests to assess the naturalness and perceived level of emphasis. Additionally, in order to gain insights into how emphasis is modelled, we evaluate the changes in prosody-related features when applying emphasis.

As already indicated in Sec. 3.1.2, we use $\text{med}_{\text{nonemph}}$ as default emphasis score and compute the score for emphasized words according to (1). For the following evaluation, α is set to 0.0, 0.5, 1.0, or 1.5, respectively, to define the four levels of increasing emphasis "input level 0" to "input level 3". Please note that the model is trained with continuous emphasis scores, and we defined fixed input emphasis levels only for the purpose of the following experiments. Since $\mathcal{M}_{\text{DD}}$ realizes emphasis by scaling phoneme durations, we multiply the durations of the emphasized word by 1.0, 1.25, 1.5, or 1.75 for "input level 0" to "input level 3," respectively. These levels are referred to as "input levels" because we use them to control the emphasis in the synthesized speech. The relation of these input levels to the perceived emphasis levels ("not emphasized", "slightly emphasized", "emphasized", "strongly emphasized") is evaluated in a listening test (see Sec. 3.2.2). We randomly select 11 test sentences from DB-E (three sentences for three of the speakers and two sentences for the remaining speaker). These sentences are synthesized using $\mathcal{M}_{\text{RNN}}$, $\mathcal{M}_{\text{XLSR}}$, $\mathcal{M}_{\text{RNN+XLSR}}$, $\mathcal{M}_{\text{CNN}}$, $\mathcal{M}_{\text{DD}}$, $\mathcal{M}_{\text{Anchor}}$, and $\mathcal{M}_{\text{NoEmph}}$. Where applicable, we generate speech samples with varying emphasis input levels, where the words to be emphasized are determined in advance. As speaker identities, we use "female1," "female2," "male1," and "male2" (see Table 1). The listening tests are conducted using the webMUSHRA [23] framework.

3.2.1. Subjective Evaluation of Naturalness

We conducted a MUSHRA-like multi-stimulus test with hidden reference and anchor [24]. As reference, we used the recordings of the test sentences from DB-E. The test consisted of 11 pages, corresponding to the 11 test sentences. In order to evaluate the impact of the emphasis levels on the naturalness, we include the synthesized audio stimuli with varying emphasis levels. On each page, the synthetic speech stimuli generated by all approaches were presented together with the reference. For each participant, we chose a random speaker and emphasis input level for each test sentence. Participants could rate the naturalness of the audio samples on a scale from 0 to 100. They were informed about which word might be emphasized. However, the presence or absence of emphasis on this specific word should not influence their rating, i.e., the participants should not rate whether it is natural to emphasize a certain word, and whether the emphasis level is perceived as natural. An open reference was available, and the participants were instructed to rate the corresponding hidden reference with a score of 100. The test was completed by 15 participants. We excluded one participant due to consistent low ratings for the reference signal. Fig. 3 depicts the results of the listening test. Comparing the naturalness scores from $\mathcal{M}_{\text{RNN}}$, $\mathcal{M}_{\text{XLSR}}$, and $\mathcal{M}_{\text{RNN+XLSR}}$ shows that these models with controllable emphasis achieve comparable or higher scores than the model without emphasis control for all emphasis input levels between "input level 1" and "input level 3". The models $\mathcal{M}_{\text{RNN}}$ and $\mathcal{M}_{\text{RNN+XLSR}}$ are slightly less natural for "input level 0". Model $\mathcal{M}_{\text{CNN}}$ received lower ratings, and model $\mathcal{M}_{\text{DD}}$ was perceived as unnatural for strong emphasis ("input level 3").

To further evaluate whether adding controllable emphasis has an impact on the overall naturalness of a model, we compare the scores of the approaches $\mathcal{M}_{\text{RNN}}$, $\mathcal{M}_{\text{XLSR}}$, and $\mathcal{M}_{\text{RNN+XLSR}}$, $\mathcal{M}_{\text{CNN}}$, and $\mathcal{M}_{\text{DD}}$, together with $\mathcal{M}_{\text{Anchor}}$, $\mathcal{M}_{\text{NoEmph}}$, and the reference in a pairwise Wilcoxon signed-rank test (with p-level threshold $p < 0.001$). In the test, the scores for all input emphasis levels are grouped together, where applicable. After adjusting the p-values with the Bonferroni method, the results show that (as expected) $\mathcal{M}_{\text{Anchor}}$ was significantly lower rated, and the reference was significantly higher rated compared to all other approaches. No significant differences are found between $\mathcal{M}_{\text{NoEmph}}$ and the evaluated approaches $\mathcal{M}_{\text{RNN}}$, $\mathcal{M}_{\text{XLSR}}$, $\mathcal{M}_{\text{RNN+XLSR}}$, $\mathcal{M}_{\text{CNN}}$, and $\mathcal{M}_{\text{DD}}$. This indicates that introducing controllable emphasis does not come at the expense of lower naturalness.

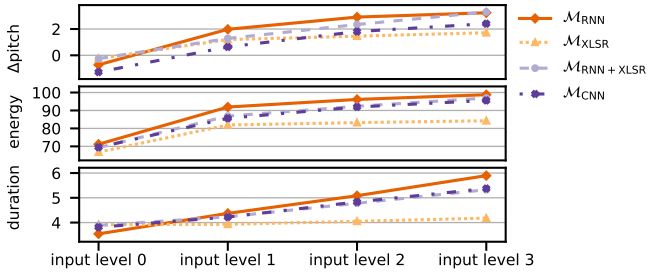

Figure 5: *Pitch, energy, and duration of emphasized words depending on the input emphasis level. The pitch is given in distance to the mean pitch of a speaker in semitones. Energy is given as the L2-norm of the mel spectrogram frames and duration refers to the number of mel frames belonging to a phoneme.*

3.2.2. Subjective Evaluation of Perceived Emphasis

We also conducted a Mean Opinion Score test using the 11 test sentences. For each audio stimulus, we randomly selected a speaker identity and emphasis input level. The participants received a single stimulus at a time together with the information which word could be emphasized. The task was to rate the perceived level of emphasis for this specific word on a likert scale with options (1) "not emphasized," (2) "slightly emphasized," (3) "emphasized," and (4) "strongly emphasized." The test was completed by 17 participants. Fig. 4 shows that especially for $\mathcal{M}_{\text{RNN}}$, but also for $\mathcal{M}_{\text{RNN+XLSR}}$ and $\mathcal{M}_{\text{CNN}}$, the perceived emphasis increases with increased input emphasis levels. We measure the relation of the input emphasis level to the perceived emphasis by computing the Pearson correlation coefficients and corresponding p-values (see Tab. 3). A Pearson correlation coefficient of 0 would indicate that there is no correlation and a coefficient of 1 would indicate that the perceived emphasis increases exactly linear with increased input emphasis. Together with the low p-values reported in the table, we can assume that the strongest correlation is achieved with $\mathcal{M}_{\text{RNN}}$, followed by $\mathcal{M}_{\text{RNN+XLSR}}$ and $\mathcal{M}_{\text{CNN}}$. Since a high correlation does not imply that there is also a high difference between perceived emphasis levels depending on the input level, we fit linear least-squares regression lines to the perceived emphasis scores of the models and also report their slopes in Tab. 3. In our experiments, the highest slopes are obtained by the same models which also achieved the highest correlation.

3.2.3. Objective Evaluation of the Effect of Input Emphasis Levels on Prosody Features

To evaluate if the prosody features change reasonably depending on the input emphasis levels, we depict the mean of the predicted features for the emphasized words in our evaluation examples in Fig. 5. Since pitch ranges vastly differ between the speakers, we use the distance to the mean pitch of the respective speaker in semitones. It can be observed that all prosody-related features increase with higher input emphasis levels. With "input level 3," we show that the model is able to generalize to emphasis levels outside the value range seen during training by raising pitch, energy, and duration in a similar fashion. Furthermore, we can now relate the perceived emphasis (Fig. 4) to the increase in prosody feature values (Fig. 5). Our experiments show that $\mathcal{M}_{\text{RNN}}$ achieves the highest perceived emphasis for input level 3 compared to other models, while also strongly changing the prosody features with increased emphasis. In contrast, the prosody features are only slightly changed with higher input emphasis levels when using $\mathcal{M}_{\text{XLSR}}$, which explains the lower ratings for the perceived emphasis levels.

4. Conclusion

Emphasis is an important factor for conveying information in speech. Therefore, it is desirable to extend speech synthesis systems to include emphasis control. This work shows that it is possible to train an emphasis detection model on a rather small publicly available dataset containing English and French recordings. We demonstrate that this model, as well as other available emphasis detection models, can be used to generate scores for training English TTS systems with controllable emphasis. The trained TTS models achieve a high naturalness according to our subjective evaluation. Furthermore, we show that increased input emphasis levels correspond to an increased perception of emphasis in the synthesized speech, which can be explained by increased prosody features.

5. Acknowledgements

This research was partially supported by the Free State of Bavaria in the DSAI project and by the Fraunhofer-Zukunftsstiftung. 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). The hardware is funded by the German Research Foundation (DFG). We thank all participants of our listening tests.

6. References

[1] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," *Language and Cognitive Processes*, vol. 25, pp. 1044–1098, 2010.

[2] M. de Seyssel, A. D'Avirro, A. Williams, and E. Dupoux, "EmphAssess : A prosodic benchmark on assessing emphasis transfer in speech-to-speech models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, 2024, pp. 495–507.

[3] M. Morrison, P. Pawar, N. Pruyne, J. Cole, and B. Pardo, "Crowdsourced and automatic speech prominence estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seoul, Republic of Korea, 2024, pp. 12 281–12 285.

[4] A. Joly, M. Nicolis, E. Peterova, A. Lombardi, A. Abbas, A. van Korlaar, A. Hussain, P. Sharma, A. Moinet, M. Lajszczak, P. Karanasou, A. Bonafonte, T. Drugman, and E. Sokolova, "Controllable emphasis with zero data for text-to-speech," in *Proceedings of the ISCA Workshop on Speech Synthesis (SSW)*, Grenoble, France, 2023, pp. 113–119.

[5] S. Seshadri, T. Raitio, D. Castellani, and J. Li, "Emphasis control for parallel neural TTS," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Incheon, Korea, 2022, pp. 3378–3382.

[6] F. Niu and W. Silamu, "Prosody-enhanced Mandarin text-to-speech system," in *International Conference on Advances in Computer Technology, Information Science and Communication (CTISC)*, Shanghai, China, 2021, pp. 67–71.

[7] Y. Zhong, C. Zhang, X. Liu, C. Sun, W. Deng, H. Hu, and Z. Sun, "EE-TTS: Emphatic expressive TTS with linguistic information," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Dublin, Ireland, 2023, pp. 4873–4877.

[8] Z. Malisz, H. Berthelsen, J. Beskow, and J. Gustafson, "PROMIS: A statistical-parametric speech synthesis system with prominence control via a prominence network," in *Proceedings of the ISCA Workshop on Speech Synthesis (SSW)*, Vienna, Austria, 2019, pp. 257–262.

[9] L. Liu, J. Hu, Z. Wu, S. Yang, S. Yang, J. Jia, and H. Meng, "Controllable emphatic speech synthesis based on forward attention for expressive speech synthesis," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, 2021, pp. 410–414.

[10] A. Suni, S. Kakouros, M. Vainio, and J. Šimko, "Prosodic prominence and boundaries in sequence-to-sequence speech synthesis," in *Proceedings of the International Conference on Speech Prosody*, Tokyo, Japan, 2020, pp. 940–944.

[11] T. Raitio, J. Li, and S. Seshadri, "Hierarchical prosody modeling and control in non-autoregressive parallel neural TTS," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore, 2022, pp. 7587–7591.

[12] S. Shechtman, R. Fernandez, and D. Haws, "Supervised and unsupervised approaches for controlling narrow lexical focus in sequence-to-sequence speech synthesis," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, 2021, pp. 431–437.

[13] T. A. Nguyen, W.-N. Hsu, A. D'Avirro, B. Shi, I. Gat, M. Fazel-Zarani, T. Remez, J. Copet, G. Synnaeve, M. Hassid, F. Kreuk, Y. Adi, and E. Dupoux, "EXPRESSO: A benchmark and analysis of discrete expressive speech resynthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Dublin, Ireland, 2023, pp. 4823–4827.

[14] P.-E. Honnet, A. Lazaridis, P. N. Garner, and J. Yamagishi, "The SIWIS French speech synthesis database – design and recording of a high quality French database for speech," University of Edinburgh, School of Informatics, The Centre for Speech Technology Research, Tech. Rep., 2017.

[15] F. Zalkow, 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)*, Rhodes Island, Greece, 2023, pp. 1–5.

[16] J. W. Kim, 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)*, Calgary, AB, Canada, 2018, pp. 161–165.

[17] E. Bakhturina, 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, pp. 2776–2780.

[18] A. Bonafonte, 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)*, Genoa, Italy, 2006, pp. 311–314.

[19] F. Zalkow, 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*, Grenoble, France, 2023, pp. 63–68.

[20] A. Mustafa, 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)*, Toronto, Canada, 2021, pp. 6034–6038.

[21] F. Lux and N. T. Vu, "Language-agnostic meta-learning for low-resource text-to-speech with articulatory features," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, 2022, pp. 6858–6868.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015.

[23] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA – A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, p. 8, 2018.

[24] International Telecommunications Union, "Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," Geneva, Switzerland, recommendation, 2015.