Controlling Prosody in End-to-End TTS: A Case Study on Contrastive Focus Generation

Siddique Latif[†*] **Inyoung Kim**[‡] **Ioan Calapodescu**[‡] **Laurent Besacier**[‡]

[†]University of Southern Queensland, Australia [‡]NAVER LABS Europe

first.last@naverlabs.com

Abstract

While End-2-End Text-to-Speech (TTS) has made significant progresses over the past few years, these systems still lack intuitive user controls over prosody. For instance, generating speech with fine-grained prosody control (prosodic prominence, contextually appropriate emotions) is still an open challenge. In this paper, we investigate whether we can control prosody directly from the input text, in order to code information related to contrastive focus which emphasizes a specific word that is contrary to the presuppositions of the interlocutor. We build and share a specific dataset for this purpose and show that it allows to train a TTS system were this fine-grained prosodic feature can be correctly conveyed using control tokens. Our evaluation compares synthetic and natural utterances and shows that prosodic patterns of contrastive focus (variations of Fo, Intensity and Duration) can be learnt accurately. Such a milestone is important to allow, for example, smart speakers to be programmatically controlled in terms of output prosody.

Index Terms: End-to-End TTS, fine-grained prosody control, contrastive focus, interrogative/assertive sentences.

1 Introduction

Text-to-Speech (TTS) systems attempt to produce human-like speech by processing natural language text inputs. Neural network based TTS has made rapid progresses and attracted a strong attention in recent years (Wang et al., 2017; Shen et al., 2018b; Łańcucki, 2021; Valle et al., 2020). Not only synthetic speech quality but also inference speed were improved, the latter benefiting from non-autoregressive TTS models (Ren et al., 2019, 2020). Controlling prosody is another issue, as features such as pitch and duration are difficult to predict because of their large fluctuations over time. While recent works proposed to synthesize speech which closely resembles the prosody of a provided reference speech using latent representations (Skerry-Ryan et al., 2018; Lee and Kim, 2019; Habib et al., 2019; Sun et al., 2020b), this work addresses explicit user control of prosody using the symbolic (text) input. Our case study concerns control of contrastive focus which emphasizes on a specific word that is contrary to the presuppositions of the interlocutor (figure 1). More precisely, we want the TTS system to correctly convey the information initially transmitted by the speaker including prosodic prominence (for instance we want to emphasize the word HOUSE in the sentence *'Sarah closed the HOUSE'*).

We posit that controlling prosody in TTS is important for future prosody transfer in speech-to-speech translation systems (carrying the real meaning of a source utterance spoken by a human). For instance, spoken utterances of figure 1 in English (where focus is only prosodically marked) would lead to different word orderings if translated to a language like Hungarian (were focus is explicitely marked by putting the verb right *after* the focused word).

The contributions of this paper are the following:

- we record and release a mono-speaker corpus of 36k English utterances usable for prosody-controlled TTS as well as for phonetic analyses on contrastive focus in English,[1]
- we show that this corpus can be used for modeling contrastive focus in English TTS and thus demonstrate that controlling prosody directly from the input text is possible,
- we compare the prosodic patterns of contrastive focus from natural and synthetic speech and provide synthetic speech samples as additional multimodal material.

The rest of this article goes simply as following: section 2 presents the background in End-to-End TTS and controllable prosody synthesis; section 3 introduces the corpus recorded for prosody-controlled TTS. Section 4 and 5 present the TTS model trained and its evaluation respectively. Finally section 6 concludes this work and gives some perspectives.

*This study was conducted during the internship at NAVER LABS Europe.

[1] https://europe.naverlabs.com/research/publications/controlling-prosody-in-tts

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 544–551
November 10–11, 2021. ©2021 Association for Computational Linguistics

(a)	Q. What happened ?	A. SARAH CLOSED THE HOUSE.
(b)	Q. Did Sarah closed the house ?	A. Sarah closed the house.
(c.1)	Q. Who closed the house ?	A. SARAH closed the house.
(c.2)	Q. What did Sarah do ?	A. Sarah CLOSED the house.
(c.3)	Q. What did Sarah closed ?	A. Sarah closed the HOUSE.
(d.1)	Q. Ava closed the house?	A. SARAH closed the house.
(d.2)	Q. Sarah occupies the house?	A. Sarah CLOSED the house.
(d.3)	Q. Sarah closed the parking?	A. Sarah closed the HOUSE.

Figure 1: Example of four focus types.; (a) broad focus, (b) given (no focus), (c) narrow focus on (c.1) subject, (c.2) verb and (c.3) object, and (d) contrastive focus on (d.1) subject, (d.2) verb, (d.3) object. We address contrastive focus here.

2 Background

2.1 End-to-End TTS Models

While former TTS systems involved complex pipelines of components optimized independently (Taylor, 2009), end-to-end neural TTS architectures were recently introduced. Such integrated models can be trained on <text, audio> pairs with minimal human annotation. Among those systems, autoregressive models (e.g., Tacotron (Wang et al., 2017; Shen et al., 2018b) and Deep voice 3 (Ping et al., 2018)) suffer from slow inference speed and robustness. Recently, non-autoregressive TTS models including Fastspeech (Ren et al., 2019, 2020), Fastpitch (Łańcucki, 2021), and JDI-T (Lim et al., 2020) address these issues by generating mel-spectrograms with extremely fast speed, while achieving comparable voice quality with previous autoregressive models. In this paper, we will use Fastpitch which is a fully-parallel TTS system, conditioned on pitch contours. It enables faster synthesis of mel-spectrograms (over 60× faster than real-time) and achieve better mean opinion scores (MOS) compared to Tacotron 2.

2.2 Related work on controllable neural prosody synthesis

Pioneering works on paralinguistic translation in speech-to-speech systems (relying on HMM-based ASR and TTS systems) (Anumanchipalli et al., 2012; Aguero et al., 2006; Tsiartas et al., 2013) were only limited to F0 and did not consider other acoustic features such as duration or intensity. Most recent works (Skerry-Ryan et al., 2018; Lee and Kim, 2019; Habib et al., 2019; Sun et al., 2020b) focused on synthesizing speech which closely resembles the prosody of a provided reference speech and try to control local prosody by varying the values of corresponding latent features. For instance, Lee and Kim (2019) introduce a fine-grained structure to encode the prosody associated with each phoneme in the input sequence using a latent variable model. Sun et al. (2020a) extends this work by incorporating a quantized latent representation to avoid discontinuous and unnatural artifacts induced by the initial approach of Lee and Kim (2019). The same authors (Sun et al., 2020b) augment Tacotron2 (Shen et al., 2018a) with a hierarchical latent variable model (at utterance, word and phone level). Finally, Zhu and Xue (2020) propose an embedding vector to continuously control the emotion strength in a TTS system.

All these methods aim at modifying prosodic attributes (duration, f0, energy) without affecting speaker characteristics but they do not provide explicit control of prosody from the symbolic (text) input. More related to our work is the approach of Wang et al. (2018) which introduces “global style tokens” (GSTs), a set of embeddings which can be seen as soft interpretable labels used to control TTS (speed, speaking style). This approach is also conceptually related to the work of Sun et al. (2020a) since it learns a quantized representation of its input. However, those style tokens are difficult to interpret since they represent meaningless prosodic dimensions learnt in an unsupervised way. Another line of work (Morrison et al., 2020) incorporates explicit user control into a prosody generation model but this is done through manually modifying the prosodic attributes such as f0 contour.

Our work shows that explicit control can be made from the symbolic (text) input: at word level (focus), utterance level (affirmative or interrogative form) or both through composition. To our knowledge, TTS generation of contrastive focus was not addressed yet in End-to-End TTS, however Pitrelli et al. (2006) did produce contrastive emphasis using concatenative TTS and Do et al. (2017) did propose a module which synthesizes emphasized speech using HMM-based TTS. We were also very recently aware of this work (Shechtman et al., 2021) that is contemporary to ours.

3 Corpus Creation

3.1 Contrastive focus

Prosody includes rhythm, pause, loudness and melody, and it reflects not only speaker’s personal state like emotion, but also linguistic information like syntax, semantics and pragmatics (among many see (Ladd, 2008)). In English, an identical text can be spoken prosodically different. In other words, prosodic prominence on different lexical units conveys different meanings. This prosodic prominence is observed on focused words in a question-answer dialogue and corrective contrasts are realised through prosodic prominence (Rooth, 1995; Büring, 2012). Figure 1 shows the four types of focus. The first 'broad focus' type is the case when the whole statement is the answer, and all words are focused (a), the second type, 'given', is when there is no focus, as all information in the statement is given (b). The third type, 'narrow focus', is when a new information is given by answering to a wh-question (c). In this case, the focus should be on the precise word in the answer according to the question, i.e. subject, verb, or object. The fourth type, 'contrastive focus', is when the answer corrects the information in the question, and the word is prosodically emphasized to convey this focused information. Roettger et al. (2019) showed the 4 distinctive F0 pitch forms corresponding to each of 4 focus types in human speech production, and we chose to address contrastive focus, as this category shows the most prominent form regarding the pitch of focused word.

3.2 Defining text prompts

Starting from a seed of 50 short sentences, similar to what is presented in Figure 1(d), we expanded them using BERT (Devlin et al., 2019), which stands for Bidirectional Encoder Representations from Transformers. We used Hugging Face (Wolf et al., 2019) library for this task. We simply masked the different subject, verb and object words in our initial utterances and let BERT predict the masked words. More than 10K utterances were generated this way in order to expand our initial sentence set. The full corpus was then manually verified before recording the corresponding speech. We removed the sentences which were not semantically correct. After manual validation of the full corpus, we kept 7320 sentences.

3.3 Recording

A professional American English female speaker was recruited for the human speech recording. The recording from prompted utterances was done remotely by the speaker from her home. After a short training session, she recorded 732 sessions where each session contained 10 groups of target sentences and each group contained five versions of the same utterance: neutral (declarative), question (interrogative), contrastive focus on subject, contrastive focus on verb and contrastive focus on object (see figure 1(d)). The neutral and question sentences were presented with a corresponding punctuation, a full stop and a question mark respectively. To elicit contrastive focus, a question-answer pair was prompted: the question in text was presented on the first line on the screen and the answer statement to this question with the focused word in upper case was presented to be read as a target sentence.

The recordings were saved on our server as audio and text pairs. We performed some postprocessing to filter out problematic samples. Finally, we got 26.7 hours of recorded signal for a total of 36600 recorded utterances. The complete corpus is made publicly available. A few sound samples are provided in the supplementary material file associated to this paper.

4 End-to-End TTS System

4.1 Model used

We use Fastpitch (Łańcucki, 2021) for experiments which is a Transformer based TTS model conditioned on fundamental frequency contours which produces state-of-the-art results. Its architecture is based on Fastspeech (Ren et al., 2019, 2020), which is composed of two Feed-Forward Transformer (FFTr) stacks. The first FFTr produces a hidden representation h from the input sequence, which is used for duration and average pitch prediction for every character using a duration and a pitch predictor module. The sum of the pitch embedding and the hidden representation h is given to the second FFTr to produce the Mel-spectrogram. WaveGlow (Prenger et al., 2019) is used as vocoder to generate English speech signal from the Mel-spectrograms.

4.2 Model training

For training the audio signals are sampled at 22KHz and silences at begin and end of the utterances are trimmed with a threshold of 30dB. Similarly to machine translation in case of domain adaptation (Kobus et al., 2016) or inline casing (Bérard et al., 2019), we annotate the input text with various control tags to mark specific prosodic elements, as shown in Figure 2. For the interrogative and declarative case, global tags at the sentence level are used to distinguish question and neutral. For the contrastive focus, we insert local tags, at the word level, just prior to the word to be focused.

NVIDIA FastPitch implementation[2] is used with the default training parameters. Each FFTr consists of a 1-D conv with ReLU activation followed by dropout and layer norm. Both duration and pitch predictors have same architecture: 1-D conv layers with ReLU, layer norm and dropout layers. Dropout rate of 0.1 is used. LAMB optimizer (You et al., 2019) is used with learning rate 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1^{-9}$.

We split the data into train, valid, and test (80/10/10 random split based on full groups of 5 sentences) and train the model for 1000 epochs. For decoding, a preexisting WaveGlow (Prenger et al., 2019) model trained on LJSpeech (Ito and Johnson, 2017) is used as vocoder. Training the model on more data using multi-speaker end-to-end TTS approaches is left for future work.

5 Evaluation of Synthetic Contrastive Focus

5.1 Natural and synthetic samples

Natural and synthetic samples are provided in supplementary material. More precisely we provide 4 groups

[2] https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/SpeechSynthesis/FastPitch

(a)	neutral	sarah closed the house
(b)	question	<Q> sarah closed the house
(c)	focus subject	<F> sarah closed the house
(d)	focus verb	sarah <F> closed the house
(e)	focus object	sarah closed the <F> house

Figure 2: For training the input text was stripped of punctuation, lowercased and annotated with global control tags <Q> for questions and local control tags <F> for focused terms. Neutral sentences have no dedicated tag.

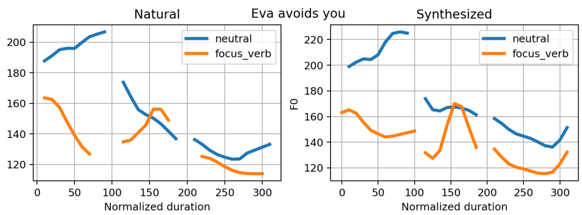
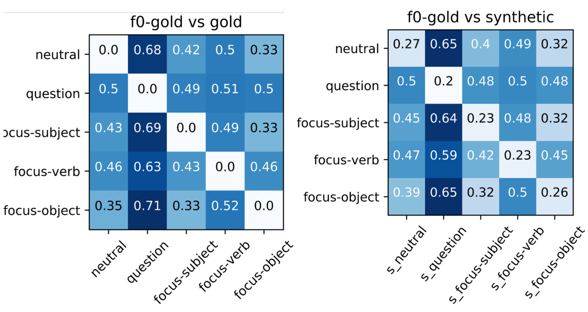


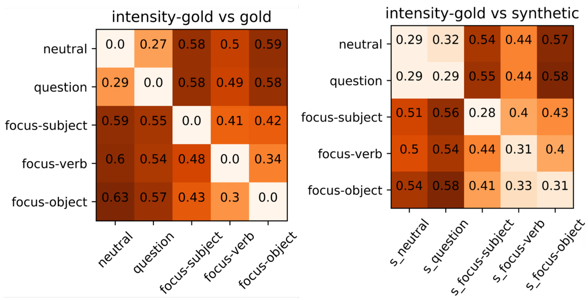
Figure 4: F0 plot of 'Eva avoids you'; Natural human voice (left) and synthesized voice from the model (right)

of wav files (from validation set) which are the recordings from female speaker where each group contains 5 sentence types (neutral, question, focus_subject, focus_verb and focus_object). We also provide the synthetic counterpart of those utterances, obtained with our TTS model learnt using the train set of our corpus. Listening to those examples show that our trained TTS model is able to convey prosodic information related to focus, a deeper quantitative analysis is proposed in the next subsection. The complete natural speech corpus (36k utterances) is also shared with the research community.

5.2 Synthetic Speech Analysis



(a) F0 evaluation



(b) Intensity evaluation

Figure 3: F0 (a) and Intensity (b) curves of each of the 5 instances (neutral, question, focus_subject, focus_verb, focus_object) are compared to each other pairwise using DTW distance. We display the confusion matrices based on those distances for natural speech (left: distance between a natural speech instance and another natural speech instance) and synthetic speech (right: distance between a synthetic speech instance and a reference natural speech instance).

For a quantitative and qualitative evaluation of our model we analyse the F0, Intensity and Duration of the generated samples. We use Dynamic Time Warping (DTW) (Sakoe and Chiba, 1978) to measure a distance directly between the F0 curves (resp. intensity curves), for each group of 5 sentences. DTW is preferred here for its ability to compute distances between temporal series of different lengths. The distance is measured between each sample type (neutral, question, focus_subject, focus_verb, focus_object) inside the gold standard (natural speech) and between the gold standard and the synthetic speech. In other words, we measure the variation of the prosodic patterns within natural speech (gold *vs* gold) to highlight the differences between focused and neutral sentences (or between different places of focus), as well as between natural and synthetic speech (gold *vs* synthetic) to evaluate the ability of our TTS model to reproduce prosodic patterns correctly. The results are averaged on our test corpus (590 groups of 5 sentences, corresponding to 2950 audio files).

We summarize the results in Figure 3 as heatmaps to visualize at once the relative distances between each feature type in the natural and synthetic speech. For each prosodic feature (F0 3a, Intensity 3b), the left matrix compares the natural speech to itself and the right matrix compares the natural speech to the synthetic speech. The lines correspond the the natural speech, the columns correspond to the natural speech (left matrix) or to the synthetic speech (right matrix).

5.2.1 F0 Analysis

As seen in 3a left matrix, the natural speech shows very distinct F0 curve for each instance type. In terms of difference, the prosodic curve for questions seems the most different from the others (neutral and focused). Between the 3 types of focus the contrastive focus on verbs is the most distinct when compared to others in terms of F0 (see also the curve of a single utterance in figure 4). As shown in 3a right matrix, the synthetic speech seems to replicate relatively well these patterns. The diagonal shows a clear similarity of the pitch curves between the natural and synthetic speech. The overall patterns are also easily identifiable: synthetic questions are the most different and between contrastive focus types the verb focus seems the most distinct.

For finer-grained analysis, pitch curves in natural and synthetic voice are plotted in figure 4. F0 pitch curve differences between two sentence types (neutral and focused_verb) in natural voice are also observed in the synthesized voice generated from our model. In natural speech, focused words are reported to realize a pitch accent or a rise-falling F0 movement (Roettger et al., 2019; Ladd, 2008), and we observe similar trend for the verb ('avoids').

5.2.2 Intensity Analysis

As for the F0 curve, we can use Figure 3b to analyse the Intensity in dB of the natural and synthetic speech utterances. In terms of natural speech the neutral utterances and questions seem to be the most similar in terms of intensity curves and the focused sentences are clearly different from the other two. As shown in the right matrix our system tends to replicate the same overall patterns with clear differences in terms of intensity between neutral and questions vs focused.

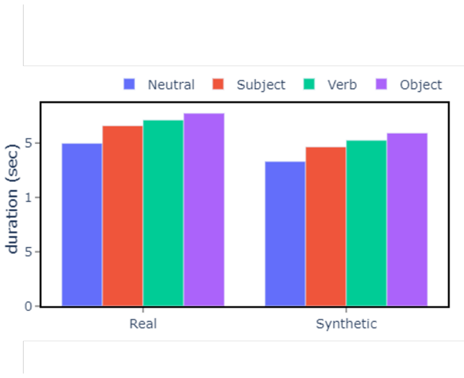
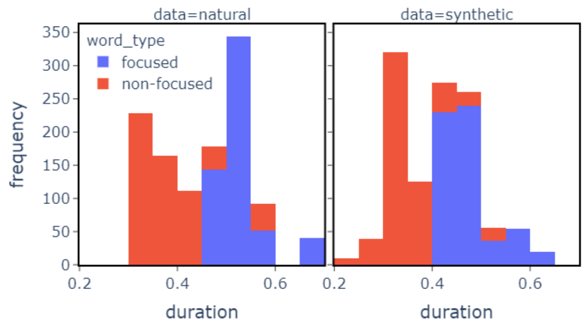


Figure 5: Comparing the duration of sentences in both real and synthetic speech for neutral and focused sentences for subject, verb, and object cases.

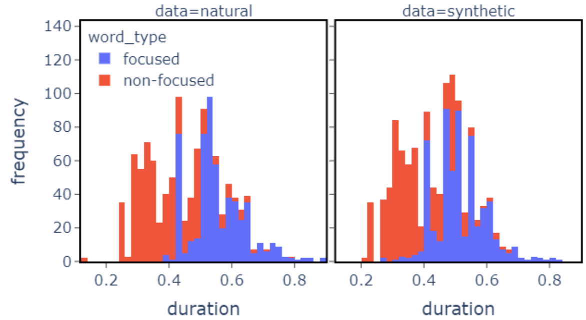
5.2.3 Duration Analysis

In Figure 5, we plot the average duration of complete sentences in both real and synthetic speech for neutral, subject, verb, and object focused cases. We observe that overall duration of focused sentences is longer than the neutral sentences both in real and synthetic speech (which displays however shorter utterances in general). It could be due to the fact that natural utterances may contain silences at the beginning and end of the recordings.

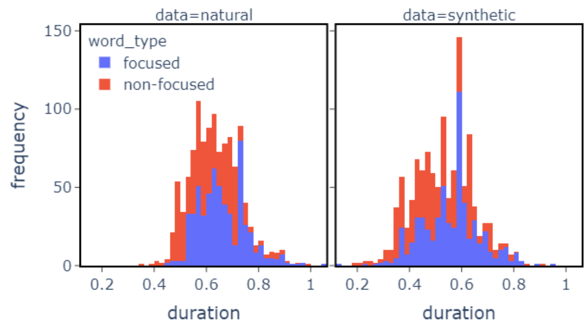
For more fine-grained analysis, we use Montreal Forced Aligner (McAuliffe et al., 2017) and compute duration of words in both (natural and synthetic) validation sets (2950 utt). We plot distributions of focused words for subject, verb, and object and compare them to those of corresponding non-focused words in Figure 6. We have less unique words for focus on subject therefore our distributions have wider bars in that case. Results on natural speech confirm that speaker reliably marked focus location (subject, verb, or object) using longer duration, as observed in Breen et al. (2010). The



(a) Subject focus vs non focus



(b) Verb focus vs non focus



(c) object focus vs non focus

Figure 6: Distribution of word durations in natural (left) and synthetic (right) speech for utterances were contrastive focus is put on (a) subject, (b) verb and (c) object.

duration difference is however less distinct when the focus is put on the object. Results on synthetic speech display similar contrast between focused and non focused words which demonstrate our TTS model has learnt to control duration in synthetic speech in order to put focus on a given word.

Statistical analysis of these distributions is performed with 2 samples Kolmogorov-Smirnov test (Massey Jr, 1951) using alpha=0.05. We plot statistics D and critical value c in Figure 7: for distributions to be identical, D should be less than c. Figure 7 shows that D values are greater than c and that non-focus and focus duration distributions are not identical for natural speech. Simililar trend (and similar D levels) is observed for synthetic speech.

Finally, we further analysed neutral and focused sentences by inspecting the *pauses* in between words

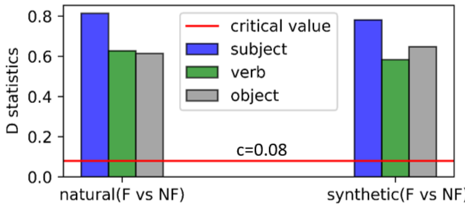


Figure 7: Illustration of statistical difference between word duration distributions for non-focused and focused words, using 2 samples Kolmogorov-Smirnov test. Left: natural speech. Right: synthetic speech.

(we ignored the starting and ending pauses in all sentences). The detection of pauses is given by the Montreal Forced Aligner. For subject we count the pauses made *after* the focused word; for object we count the pauses made *before* the focused word; and for verb we count the pauses made *before* and *after* the focused word. Table 1 compares those counts for focused or neutral sentences for both natural and synthetic speech. We observe that focused sentences have more detected pauses compared to the neutral ones and that our TTS model is also able to reproduce this pattern in synthetic speech.

Table 1: Comparing the number of detected pauses between focused and neutral sentences, for both natural and synthetic speech.

Number of pauses in natural speech around Subj/Verb/Obj words					
Subject focused	Subject neutral	Verb focused	Verb neutral	Object focused	Object neutral
425	3	446	32	488	33
Number of pauses in synthetic speech around Subj/Verb/Obj words					
502	2	472	43	516	41

5.3 Varying amount of training data

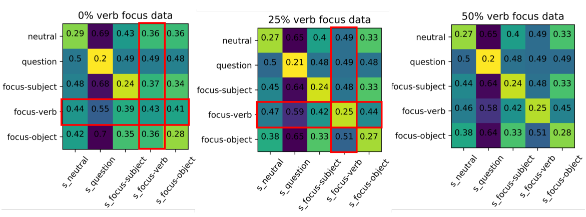


Figure 8: Varying the amount of training data containing utterances with focus_verb: 0%, 25% and 50%. We display only DTW distances between F0 curves of each of the 5 instances (neutral, question, focus_subject, focus_verb, focus_object) compared to each other (natural *vs* synthetic).

To test the generalisation capacity of the model, we trained with three alternative subsets of the data with respectively 0% of training data for focus_verb, 25% and 50%. These variations should be able to tell us if the model is able to generate focused words in positions it has not seen before (here verb) and if not how much data is needed for the model to generalize? Results for the F0 curves comparison are shown in Figure 8 (trend is similar for intensity and duration). As we see in the first heatmap, without any training utterances containing focus_verb, the model has difficulties to generalize (see 4th line and column). The model is able to clearly differentiate with the questions but the fine grained differences between focus words of different positions were not captured well. This is confirmed by a manual inspection we did on some samples: the model produced sometimes neutral sounding sentences, sometimes focus was misplaced on another word and sometimes it was correct. When we slightly increase the amount of utterances containing this event (25%) the model is able to regain good performance and displays the same similarity patterns as the ones we observed with the full training set.

6 Conclusion and Future work

In this paper, we proposed a control mechanism for End-to-End TTS systems to manipulate fine-grained prosodic features like contrastive focus (as well as affirmative vs interrogative sentences). This mechanism uses local and global interpretable control tags directly in the input sequences to manipulate the generated prosody: F0, Intensity, pauses and durations. A specific mono-speaker corpus was recorded for this study and we hope its release will help the speech community to (a) study fine grain prosodic patterns of focus that occur in natural speech, (b) continue investigating the problem of fine grained control of prosody for TTS models and (c) address the dual problem of automatically detecting prosodic focus from natural speech.

In future work we plan to extend this study with a stronger TTS baseline and human evaluations. We would also like to verify how to train a multispeaker TTS system using this same dataset. It would allow to create a generic TTS model with, as base, a non annotated dataset and, as extensions, many smaller corpora with specific prosodic annotations. Naturally the extension of this work to the complementary task seems also straightforward (i.e. building an ASR system with automatic prosodic feature extraction) as well as the concrete application of the technology to Speech-to-speech Neural Machine translation with prosodic features transfer between the input signal and the output translation, as was very recently initiated by (Tokuyama et al., 2021). Finally, the possibility to emphasize not only word but also subword units is another interesting perspective.

7 Acknowledgements

We thank Jennifer, our American English speaker, for her professional speech recordings made following our precise instructions. We also thank the anonymous reviewers for their comments and suggestions on our work.

References

P.D. Aguero, J. Adell, and A. Bonafonte. 2006. Prosody generation for speech-to-speech translation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

Gopala Krishna Anumanchipalli, Luís C. Oliveira, and Alan W Black. 2012. Intent transfer in speech-to-speech machine translation. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 153–158.

Mara Breen, Evelina Fedorenko, Michael Wagner, and Edward Gibson. 2010. Acoustic correlates of information structure. *Language and Cognitive Processes - LANG COGNITIVE PROCESS*, 25:1044–1098.

Daniel Büring. 2012. Focus and intonation.

Alexandre Bérard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs europe's systems for the wmt19 machine translation robustness task.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Quoc Truong Do, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura. 2017. Preserving word-level emphasis in speech-to-speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 25(3):544–556.

Raza Habib, Soroosh Mariooryad, Matt Shannon, Eric Battenberg, RJ Skerry-Ryan, Daisy Stanton, David Kao, and Tom Bagby. 2019. Semi-supervised generative modeling for controllable speech synthesis. In *International Conference on Learning Representations*.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Catherine Kobus, Josep Maria Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *CoRR*, abs/1612.06140.

D. Robert Ladd. 2008. *Intonational Phonology*, 2 edition. Cambridge Studies in Linguistics. Cambridge University Press.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. *ICASSP*.

Younggun Lee and Taesu Kim. 2019. Robust and fine-grained prosody control of end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5911–5915. IEEE.

Dan Lim, Won Jang, O Gyeonghwan, Heayoung Park, Bongwan Kim, and Jaesam Yoon. 2020. Jdi-t: Jointly trained duration informed transformer for text-to-speech without explicit alignment. *Proc. Interspeech 2020*, pages 4004–4008.

Frank J Massey Jr. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proc. Interspeech 2017*, pages 498–502.

Max Morrison, Zeyu Jin, Justin Salamon, Nicholas J. Bryan, and Gautham J. Mysore. 2020. Controllable neural prosody synthesis. In *Interspeech 2020, Shanghai, China, 25-29 October 2020*, pages 4437–4441. ISCA.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2018. Deep voice 3: 2000-speaker neural text-to-speech. *Proc. ICLR*, pages 214–217.

J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny. 2006. The ibm expressive text-to-speech synthesis system for american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1099–1108.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.

Yi Ren, Chenxu Hu, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text-to-speech. *arXiv preprint arXiv:2006.04558*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*.

T. Roettger, Tim Mahrt, and Jennifer Cole. 2019. Mapping prosody onto meaning – the case of information structure in american english*. *Language, Cognition and Neuroscience*, 34:841 – 860.

Mats Rooth. 1995. A theory of focus interpretation. *Natural Language Semantics*, 1.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49.

Slava Shechtman, Raul Fernandez, and David Haws. 2021. Supervised and unsupervised approaches for controlling narrow lexical focus in sequence-to-sequence speech synthesis. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pages 431–437. IEEE.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018a. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4779–4783. IEEE.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018b. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR.

Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran, and Yonghui Wu. 2020a. Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 6699–6703. IEEE.

Guangzhi Sun, Yu Zhang, Ron J Weiss, Yuan Cao, Heiga Zen, and Yonghui Wu. 2020b. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6264–6268. IEEE.

Paul Taylor. 2009. *Text-to-speech synthesis*. Cambridge university press.

Hirotaka Tokuyama, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Transcribing Paralinguistic Acoustic Cues to Target Language Text in Transformer-Based Speech-to-Text Translation. In *Proc. Interspeech 2021*, pages 2262–2266.

Andreas Tsiartas, Panayiotis Georgiou, and Shrikanth S. Narayanan. 2013. Toward transfer of acoustic cues of emphasis across languages. In *Proceedings of InterSpeech*.

Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro. 2020. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. *arXiv preprint arXiv:2005.05957*.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *Proc. Interspeech 2017*, pages 4006–4010.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*.

Xiaolian Zhu and Liumeng Xue. 2020. Building a controllable expressive speech synthesis system with multiple emotion strengths. *Cognitive Systems Research*, 59:151–159.