Invert-Classify: Recovering Discrete Prosody Inputs for Text-To-Speech

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ASRU57964.2023.10389773

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
ABSTRACT

Modeling prosody in Text-to-Speech (TTS) is challenging due to ambiguous orthography and the high cost of annotating prosodic events. This study focuses on the modeling of contrastive focus, the emphasis of a word to contrast it to presuppositions held by an interlocutor. Modeling of contrastive focus can be done in TTS by using binary, symbolic inputs at the word level in a supervised setting. To address the absence of annotated data, we propose the Invert-Classify method, which leverages a frozen TTS model and unlabeled parallel text-speech data to recover missing contrastive focus inputs. Our approach achieves a binary F-score of up to 0.71 for contrastive focus annotation recovery, utilizing only 5-10% of annotated training data. Furthermore, subjective listening tests show that training on additional data labeled via Invert-Classify enhances overall synthesis quality, as well as providing good control and plausible-sounding contrastive focus.

Index Terms— text-to-speech synthesis, prosody modeling

1. INTRODUCTION

Supervised Text-to-Speech (TTS) aims to predict natural and suitable-sounding synthetic speech from text input. The training process for TTS models typically involves utilizing parallel data, which consists of both text and corresponding speech waveforms. However, the majority of text data is plain and lacks explicit information about various properties of speech, including prosodic aspects such as rhythm, intonation, and loudness. As a result, TTS synthesis often lacks the desired level of controllability and variance, which are crucial for enhancing contextual appropriateness and overall quality.

In order to address the challenge of ambiguity and improve the synthesis of natural and expressive speech, there are two approaches which are commonly employed. The first approach involves incorporating a reference utterance for style transfer [1, 2, 3, 4], which enables control over aspects such as emotion and style at the utterance level. This method requires a reference utterance as input and offers controllability on a broader scale. The second approach involves incorporating symbolic inputs alongside the text to provide more explicit control over prosodic events, such as Tone and Break Indices (ToBI) [5], explicit F0 or pitch inputs [6], or emotion labels [7]. By including additional text inputs that specify prosodic events, this method allows for more precise control at the phone, word, or utterance level, enhancing the overall quality and controllability of the synthesized speech.

The focus of this research is on utilizing symbolic inputs to incorporate contrastive focus in English, which involves emphasizing specific words to contrast the presuppositions of an interlocutor [8]. In English, contrastive focus is achieved by modifying acoustic correlates such as F0 (fundamental frequency), duration, and energy, to draw attention to a word that corrects contextual information [9]. In TTS, contrastive focus can be modeled at the word level, using binary annotations that indicate whether a word should have contrastive focus or not [10]. However, obtaining these annotations from human listeners can be costly.

To address the issue of obtaining annotated data, we propose a method called Invert-Classify, which automatically derives contrastive focus labels for parallel text-speech data. The details of the Invert-Classify method are further explained in Section 3.3 and Section 3.4. We demonstrate the effectiveness of Invert-Classify in conjunction with FastPitch models [11] trained on varying amounts of training data. These findings highlight the potential of improving the efficiency of TTS models that utilize symbolic inputs for prosody specification, thereby reducing the burden of manual annotation.

We explore the following research questions to investigate the efficacy of Invert-Classify:

1. How well can Invert-Classify be used to retrieve contrastive focus labels?
2. How well are listeners perceiving contrastive focus in ground truth and synthetic stimuli?
3. How much labeled data is needed for perceivable and...
4. How helpful are retrieved contrastive focus labels for modeling TTS?

2. RELATED WORK

Invert-Classify leverages inversion and cosine classification. Initially, inversion was intended as a means for visualizing the generalization of image classification models [12]. The process involves taking a frozen model trained for image classification and minimizing a cross-entropy loss with respect to a randomly initialized input image. The randomly initialized image is then updated to produce the most optimal loss for a given target class. Notably, it was observed that models with stronger generalization capabilities produced images that closely resembled the target class when subjected to inversion.

More recently, inversion has seen some success in tasks such as text-to-image, where new style embeddings can be generated via inversion to represent new style tokens, known as text inversion [13]. With text inversion, a new token embedding can be randomly initialized and then inverted for a hand selected group of images, adapting to the style of the images in that particular group. The resulting token embedding can be used for greater control and personalization.

3. METHODS

For text to speech synthesis, we are using a two tier system that is trained separately. The first tier is a transformer-based sequence to sequence model, known as FastPitch [11], that takes artpabet phone inputs and predicts corresponding frames of mel-spectrogram. FastPitch is a non-autoregressive, state of the art model, made up of two feed forward transformer stacks. During training, FastPitch also receives F0 and duration targets obtained during preprocessing of audio for explicitly modeling duration and F0. The hidden representations from the first feed forward transformer are used as inputs for F0 and duration predictor components. Predicted F0 targets are then summed with the hidden representation from the first feed forward transformer, before being decoded by the second feed forward transformer into a sequence of frames of mel-spectrogram. For the second tier, a general adversarial network, known as HiFiGAN [14], is used as a vocoder to predict waveform from mel-spectrogram inputs.

3.1. Modeling contrastive focus

Similar to previous work, we modify FastPitch to take additional input in the form of binary symbols, indicating whether a phone occurs within a word that has contrastive focus or not [10]. This means that each phone has a corresponding annotation that indicates focus with the arbitrarily assigned values of 2 for no contrastive focus or 1 for contrastive focus. The contrastive focus annotations are passed through a separate embedding layer with K entries (referred to as a symbolic embedding layer), in which the resulting embeddings V_k are summed with the embeddings of the phone inputs and the positional embeddings, as shown in Figure 1.

3.2. Supervised learning

Initially, FastPitch is trained in a supervised setting with text and contrastive focus annotations as input, by optimizing a combination of L2 losses between the predicted and ground truth mel-spectrogram, durations, and F0 values. This combination of losses is termed the fusion loss and is calculated as the mean squared error between the predicted and ground truth mel-spectrogram \( \hat{y} \) and \( y \), F0 \( \hat{p} \) and \( p \), and duration \( d \) and \( \hat{d} \). Finally, hyperparameters for scaling weights \( \alpha \) for the pitch loss and \( \gamma \) for the duration loss, allow the complete loss function to be defined as follows:

\[
L = ||\hat{y} - y||^2 + \alpha||\hat{p} - p||^2 + \gamma||\hat{d} - d||^2
\]

3.3. Inversion

After supervised learning, unannotated data that was not part of the previously used training data is selected for inversion. As stated previously, inversion is the process of performing back propagation on a frozen model and updating inputs \( I \) to minimize the loss \( L \). To minimize the loss function \( L \) with
Fig. 2. After inversion, $I_{\text{new}}$ is cosine-classified to $V_k$ basis vectors, which serve as entries to the symbolic embedding layer for encoding discrete prosody inputs (e.g., contrastive focus) during supervised training.

respect to $I$ via ADAM [15], the following equation can be employed:

$$I_{\text{new}} = I - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (2)$$

Here, $I_{\text{new}}$ represents the updated value of $I$, $\hat{m}_t$ represents the bias-corrected first moment estimate of the gradients, $\hat{v}_t$ represents the bias-corrected second moment estimate of the gradients, $\epsilon$ represents a constant to prevent division by zero, and $\eta$ denotes the learning rate.

In this case, the input variables $I$ corresponds to the contrastive focus vectors and are initialized by taking the mean of the embedding entries $V_k$ within the symbolic embedding layer for contrastive focus.

### 3.4. Cosine classification

While optimizing the symbolic input embeddings $I$, the previously learned embeddings obtained during supervised learning serve as $K$ basis vectors $V_k$. These basis vectors are used to compute the cosine similarity between the input vector, $I_{\text{new}}$, and each of the basis vectors $V_k$, using the formula:

$$\argmax_k \frac{I_{\text{new}} \cdot V_k}{\|I_{\text{new}}\| \cdot \|V_k\|} \quad (3)$$

This process of cosine classification, as shown in Figure 2, results in the assignment of the input $I_{\text{new}}$ to one of the $K$ discrete values based on cosine similarity. After cosine classification, accuracy metrics can be obtained by comparing the assigned values to the ground truth contrastive focus annotations.

### 3.5. Inductive biases

Different conditional statements have been implemented to constrain the inversion and quicken convergence. Not allowing silence to be updated and tying all embeddings that correspond to the same word to one single embedding, achieved best results for accuracy and convergence speed. Therefore, we implement these inductive biases for all models and experiments.

### 4. EXPERIMENTAL CONDITIONS

#### 4.1. Training details and hyperparameters

During supervised training of FastPitch, a batch size of 30 is used, LAMB [16] is used as an optimization algorithm, 0.2 is set as the weights $\alpha$ and $\gamma$ to be multiplied with the losses for duration and pitch, and an overall learning rate of 0.1 is used. All Fastpitch models are trained for 1000 epochs.

During inversion, inputs are optimized with a learning rate of 0.75 to minimize the same loss function used during the supervised training of FastPitch. The inversion process is performed with a batch size of 1 and spans over 600 epochs.

For HiFiGAN, the generator and discriminators are fine-tuned on the training set, which have been pretrained on the Universal data set for 2.5 million steps with a batch size of 16 and a learning rate of 0.0002. The same hyperparameters used during pretraining are used during finetuning, except that the number of update steps is 131,000. During inference, only the generator is used.

#### 4.2. Data

The data used in this study is from the Naver Prosody-Control data set [10], which features 26.7 hours of 36600 individual speech utterances with corresponding text, from an American English female speaker. The data was recorded over the span of 732 sessions, in which each session contained 10 different, minimal pair groups of different utterances. Each minimal pair group contained five versions of the same utterance spoken differently. These different versions of the utterances were neutral (i.e., declarative), questioning, contrastive focus on the subject, contrastive focus on the object, and contrastive focus on the verb. Questioning and neutral utterances were elicited by presenting the speaker with the utterance containing either a period or question mark. Utterances containing contrastive focus were presented with a context question requiring clarification and the answer meant to be spoken with the target word for contrastive focus in capital letters. A sample of a minimal pair group in writing can be seen in Table 1.
4.3. Preprocessing

All audio was originally sampled at 44.10 kHz and then down-sampled to 22.05 kHz. 422 recordings were found to have artifacts and lower sampling rates of 16 kHz and were subsequently excluded from the data set.

Due to the data being read speech, with the text transcriptions originating before the speech recordings, this meant that the transcriptions had no corresponding time alignments to specify when specific words occurred. Therefore, a single speaker, forced alignment model was trained from scratch with the Montreal Forced Alignment API, from [17]. As per the default settings, the CMU dictionary from [18] was used to convert words to Arpabet symbols. By using the CMU dictionary, a total of 39 out of vocabulary (OOV) words were encountered, which totaled to 2004 tokens.

5. EXPERIMENTS

5.1. Classification efficacy

The effect of training set size on classification accuracy is explored by removing minimal pair groupings to form subsets which are 1%, 5%, 10%, 20%, and 100% of the original training set size. Cosine classifications are chosen based on achieving the lowest L2 fusion loss for FastPitch, before accuracy metrics are calculated. Due to computation and time constraints, classification experiments were carried out on a subset of the test set, which totaled to 508 utterances.

5.2. Controllability and quality

In addition to manipulating the training set, we conducted several experiments using different models to examine the subjective effects of available annotated data on contrastive focus, controllability, and quality. The experiments involved synthesizing utterances using various approaches, including copy-synthesis, a model trained on only neutral utterances with random contrastive focus labels, and a model trained on all data with random contrastive focus labels. Additionally, a model was trained with data annotated via Invert-Classify, along with 5% of the training data containing ground truth labels. The Invert-Classify approach utilized a FastPitch model trained on 5% of the training data.

To evaluate the synthesized utterances, we conducted listening tests with 30 American participants who were native English speakers. Each model was used to synthesize 8 minimal pair groupings, resulting in a total of 32 utterances per model. However, models trained on neutral data and randomly annotated data produced 8 unique utterances with neutral target labels.

During the listening tests, participants were asked to rate each utterance on a 1-5 scale for quality based on sound (while ignoring word choice or grammar) and identify the word they believed the speaker was trying to make prominent or emphasize. The order of the utterances was randomized and fixed for all participants. When choosing which word was believed to be prominent, participants were allowed to choose ‘None of these words’ and all words in the utterance except ‘the’ if present. Practice ground truth utterances were provided as a competency test, requiring participants to answer the correct choices for contrastive focus. All participants were asked for consent and agreed to use headphones during the study. The median completion time for the listening tests was 49 minutes, and participants were compensated at minimum wage. Data from three participants was discarded due to providing partial completion and/or erroneous answers.

6. RESULTS

6.1. Classification accuracies

As seen in Table 2, binary Fscore (with contrastive focus being the positive label) begins to drop steeply below 10% of the training set size. Recall is comparatively higher than precision in all cases, meaning that inaccuracies are likely due to over-predicting contrastive focus. The high false positive rate may be affected by a lower occurrence of contrastive focus and noise within ground truth annotations, however that is unlikely to explain all of the false positives. Overall, these results indicate a model trained with only 5-10% of annotated training data can be used to automatically recover prosodic prominence with similar performance as observed with a model trained on much larger quantities of (expensive) annotated data.

<table>
<thead>
<tr>
<th>Training Set Size</th>
<th>Precision</th>
<th>Recall</th>
<th>Fscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0.42</td>
<td>0.73</td>
<td>0.54</td>
</tr>
<tr>
<td>5%</td>
<td>0.51</td>
<td>0.92</td>
<td>0.66</td>
</tr>
<tr>
<td>10%</td>
<td>0.58</td>
<td>0.93</td>
<td>0.71</td>
</tr>
<tr>
<td>20%</td>
<td>0.56</td>
<td>0.94</td>
<td>0.70</td>
</tr>
<tr>
<td>100%</td>
<td>0.61</td>
<td>0.92</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 1. Sample Minimal Pair Group from the Naver Prosody-Control Dataset

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral Question</td>
<td>Abby went home.</td>
</tr>
<tr>
<td>Cara went home?</td>
<td>Abby went home?</td>
</tr>
<tr>
<td>Abby dax home?</td>
<td>ABBY went home.</td>
</tr>
<tr>
<td>Abby went going?</td>
<td>Abby went HOME.</td>
</tr>
</tbody>
</table>
### Table 3. Mean Opinion Score Results with 95% confidence intervals calculated from 1000 iterations of bootstrapping

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Opinion Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>1.51 ± 0.15</td>
</tr>
<tr>
<td>5%</td>
<td>2.98 ± 0.19</td>
</tr>
<tr>
<td>Invert-Classify</td>
<td>3.68 ± 0.14</td>
</tr>
<tr>
<td>10%</td>
<td>3.53 ± 0.14</td>
</tr>
<tr>
<td>20%</td>
<td>3.75 ± 0.13</td>
</tr>
<tr>
<td>100%</td>
<td>3.81 ± 0.10</td>
</tr>
<tr>
<td>Copy-Synth</td>
<td>4.07 ± 0.08</td>
</tr>
<tr>
<td>Random Labels</td>
<td>3.51 ± 0.14</td>
</tr>
<tr>
<td>Neutral</td>
<td>3.82 ± 0.29</td>
</tr>
</tbody>
</table>

#### 6.2. Mean opinion scores

The Mean Opinion Score (MOS) results in Table 3 demonstrate a notable improvement in terms of quality for the model trained on 5% ground truth labels and 95% labels retrieved via Invert-Classify. As observed in the classification results, the model trained on only 1% of the training data obtains the lowest MOS. Surprisingly, the model trained with random labels performs even worse than the model trained on 20% of the data, despite having access to 100% of the unlabeled training data. This indicates that contrastive focus labels not only contribute to controlling prosody but also have a positive impact on the MOS for quality. Furthermore, the model trained on random labels also does not outperform the model trained on random labels with only neutral utterances. More training data does not equate to better MOS for quality if the training data varies in contrastive focus and annotations are absent.

#### 6.3. Target focus perception rate

In Figure 3, we observe the performance of various models in generating stimuli for target focus perception rate. Surprisingly, the model trained with just 20% of the training data stands out, achieving the highest median target focus perception rate. However, it’s worth noting that the difference in performance compared to models trained with 100% of the training data or the copy-synthesis approach is not significant. Furthermore, when we closely examine the models that generate stimuli with the highest median scores, copy-synthesis emerges with the smallest variations between its upper and lower quartiles, relative to the median. This observation underscores the capability of achieving near-optimal controllability levels with a mere 20% of the training data. It highlights the potential of leveraging a relatively small amount of annotated data for this task.

Interestingly, when we consider the models trained with only 10% and 5% of the data, they produce stimuli that exhibit only minor deficiencies of target focus perception rate, when compared to models trained with more training data. These results deviate from the objective measures reported in previous research using the same data and model [10], which had previously estimated that FastPitch might necessitate at least 20% of the training data to generate stimuli with the correct perceived target focus.

The model trained on 5% ground truth labels and 95% Invert-Classify retrieved labels performs slightly better than the model trained on 10% of the training data. However, the difference in performance is not significant. On the other hand, the model trained on only 1% of the data exhibits notably poorer results, aligning with the trends observed in both the MOS and classification results.

Models trained on neutral utterances and random labels both had neutral target labels. Interestingly, the model trained on neutral utterances performs slightly better, despite having being trained on less data. Conversely, the model trained with random labels has been exposed to a more extensive dataset with varying degrees of contrastive focus. This unexpected diversity in the training data results in generated stimuli that unexpectedly exhibit perceivable contrastive focus, even when the intended target focus is neutral. Although neither model possesses input-level control over contrastive focus, the model trained on the more limited dataset of only neutral utterances excels in generating stimuli perceived as neutral.
7. DISCUSSION

Invert-Classify proves to be a highly effective method for retrieving contrastive focus labels even with as little as 5-10% of the training data present. However, the classification accuracy experiences a significant drop below 5%. Notably, False Positives are the primary source of error for Invert-Classify, leading to an over-prediction of contrastive focus. Despite the presence of False Positives, leveraging Invert-Classify can enhance model performance in terms of perceived quality, while maintaining a consistently high level of controllability.

One limitation of this study is that the dataset used for the experiments is confined to short utterances following a specific subject-object-verb order. As such, the performance may differ on less ideal data with a lower frequency of contrastive focus. To address this, future investigations may include exploring further optimization constraints, inductive biases, and alternative loss functions that may assist in improving the inversion process.

8. CONCLUSION

In summary, Invert-Classify presents a promising approach for obtaining contrastive focus labels, allowing for enhanced TTS model performance with limited annotated data. Despite certain limitations, this research contributes valuable insights into improving contextual appropriateness and overall quality of synthesized speech in TTS systems. Future works may explore similar methods for cross-domain, multi-speaker, or different types of prosody control.

9. REFERENCES


