Are we truly modeling expressiveness? A study on expressive TTS in Brazilian Portuguese for real-life application styles

Lucas H. Ueda1, Paula D.P. Costa1, Flavio O. Simoes2, Mario U. Neto2

1Department of Computer Engineering and Automation, School of Electrical and Computer Engineering, University of Campinas, Campinas, Brazil
2Research and Development Institute CPQD, Campinas, Brazil

1156368@dac.unicamp.br, paulad@unicamp.br, simoes@cpqd.com.br, uliani@cpqd.com.br

Abstract

This paper presents a study of expressive speech synthesis applied to real-life application styles in Brazilian Portuguese. We explore the use of data with different recording conditions in state-of-the-art architectures in expressive TTS. Our results suggest that the variability of recording conditions of the same style, combined with a guided training of the latent representation space of the Reference Encoder, assists in the modeling of non-archetypal expressivities. Additionally, we propose an alternative to evaluating the model’s ability to generate expressive speech during preliminary results, based on a classifier using GeMAPS features.

Index Terms: expressive speech synthesis, tacotron2, style guided, prosodic features, gmaps

1. Introduction

The process of generating an artificial speech from a given text is named text-to-speech (TTS) synthesis. This artificial speech should correctly convey the message that was in the input text (intelligibility) and, ideally, sounding like a human (realism/naturalness) while having the correct prosody (expressiveness). The usage of recent deep neural networks architectures in TTS, neural TTS (NTTS), have established a new state of the art, being able to generate intelligible artificial speech with a naturalness close to the human voice [1, 2, 3, 4]. However, state-of-the-art NTTS models can still not synthesize realistic, expressive speech or modulate existing models for different styles. Because of this, the synthesis of expressive speech is still a challenge.

To modulate expressiveness in speech, we need to change the prosody. According to [3, 6] prosody is speech characteristics that are not associated with “what is said”, but with “how it is said”. Similarly to [7] we define prosody as the variation in speech signal that remains after accounting for variation due to phonetics, speaker identity, and channel effects. There are several parameters that can be changed to modify speech prosody, mostly related to fundamental frequency, intensity, and duration during speech. Recent approaches have proposed a global control of prosody, where a latent representation of prosody (LRP) is estimated in conjunction with NTTS models. From this latent space, these approaches showed to be able to alter the speech prosody without having to control specific acoustic parameters. Although these approaches show that modeling speech prosody without explicit specifications is feasible, such approaches do not control the expressiveness of speech itself.

Later works seek to use these architectures based on estimating the LRP to generate different speech styles. When estimating the LRP space using data with varying speech styles, the modeled space itself showed other characteristics for each of them. From the analysis of this space, such works showed that it was possible to control the speech style. However, usually, these approaches use archetypal styles of speech, clearly distinguishable from each other.

Based on the availability of lines recorded in different recording styles and conditions, this work explores the ability of state-of-the-art models to model the prosodic space (LRP). The present work investigates the robustness of current methods of expressive speech synthesis with a diversity of recording conditions and real-life application styles.

We explore different dataset configurations under different recording conditions, as well as different methodologies for estimating LRP space. Our experiments showed that Global Style Tokens [8] are not capable of generating separable prosodic spaces, while Reference Encoder [7] easily generates separable spaces. However, the space generated does not guarantee that different styles are modeled. The use of the guidance in the estimation of the LRP showed feasibility in conditioning the space to better prosodic modeling. Additionally, we propose an alternative to subjective perceptual evaluation suitable for intermediate stages of TTS development, based on an expressivity classifier that uses GeMAPS [9] set of acoustic features.

2. Related works

Three acoustic parameters are mainly related to speech prosody: fundamental frequency, intensity, and duration. Moreover, we can classify prosody in two categories: affective and augmentative [10]. The affective prosody is the expression of meaning related to emotion, mental state, and speaker attitude. On the other hand, augmentative prosody does not contain any extra information; it is used to make a verbal communication clearly by giving intonation or focus in specific parts, disambiguating one message that could be interpreted in different ways.

Silva and Barbosa [11] tried to verify how to relate prosodic acoustic features with the perception of people listening to emotional speech. Despite a good correlation of certain features, many of them did not have a significant relationship and were unstable over different emotions. The high complexity of how prosody is related to acoustic features is a big challenge when talking about controlling expressiveness in artificial speech.

In order to avoid the need of explicit annotations in prosody modeling, Skerry-Ryan and colleagues [7] proposed the Reference Encoder (RE), which consists of an additional module to the NTTS Tacotron architecture [1] that encodes the mel spectrogram of a reference audio in a lower-dimensional representation. This representation is added to the decoder input and used to control prosody at inference time. This work showed...
that the augmented Tacotron with RE can transfer prosody from reference audio to the synthesized speech. Later, Wang and colleagues [8] proposed an attention layer augmenting the Reference Encoder, where trainable variables, named Global Style Tokens (GST), are jointly estimated with the model parameters in training time. This approach shows that each token can model distinct acoustic parameters, like pitch and duration. Although these two works have shown that it is possible to use this lower-dimensional representation to control prosody in a NTTS systems, they didn’t perform any relation between these parameters and controlling a specific speech style.

Further, other works tried to model different speech styles using these lower-dimensional representations, which we are calling Latent Representation of Prosody (LRP). In [12] the GST’s were used conditioning each token to a specific emotion label. By doing that, they were able to control the synthesized speech over three different emotions by using the respective modeled token. Kwon, Jang, Ahn and Kang [13] also used the GST architecture, but instead of conditioning the tokens themselves, they studied the LRP space generated by tokens over speech samples. This work demonstrated that using the centroid’s of each emotional speech in LRP space can lead to expressive speech control in inference time. However, those approaches used a balanced internal dataset among different styles, which is not easily feasible.

Recently, Sorin, Shechtman and Hoory [14] proposed an approach where a NTTS model augmented by RE module was used to generate expressive speech. Particularly, this work showed that it is possible to transfer expressiveness among different speakers even when just one speaker’s expressive speech is available. Using a dataset consisting of speech from 3 speakers, where just one of them had a small amount of expressive data, they trained a multi-speaker Tacotron2 [2] architecture with RE module to generate the LRP space. Using PCA decomposition, they were able to select a good representation for each style to generate controlled expressive speech in inference time to all speakers. Specifically, in this project, they used real-life application speech styles called “good news” and “apology”, besides the neutral. This work is particularly important because real-life application styles are not so easily distinguishable from each other, unlike what happens when dealing with emotional speech.

As far as we know, the only work that applies NTTS approach to Brazilian Portuguese language is in [15], where several experiments were done using different NTTS architectures on a public dataset made by themselves. However, they don’t contemplate the expressive speech synthesis problem, having only neutral samples. We understand this project as the first one considering the expressive NTTS problem using Brazilian Portuguese language.

3. Technical setup

3.1. Data

Our experiments were conducted using a proprietary dataset consisting of utterances recorded by three speakers identified as: Speaker 1 (female), Speaker 2 (female), and Speaker 3 (male). Table 1 presents a summary of the dataset, which is characterized by a high volume of neutral recordings, and a smaller set of expressive speech samples. The expressive style is associated to real-life customer service applications and can be described as excited positively; we refer to this as Enthusiastic style. Also, the expressive utterances were recorded only by Speaker 1 in two different conditions: medium quality (with high reverberation) and studio quality.

In order to create different experimental setups, four dataset configurations were designed, each one characterized by a particular selection of utterances / recording conditions. The first configuration, named DC1, consists of all available neutral data from all three speakers, together with all expressive utterances recorded in high reverberation condition. This configuration is, theoretically, characterized by a more easily separable latent subspace, since utterances differ both in terms of expressiveness and recording condition at the same time.

The other three settings are more challenging and have been used to cover three possible combinations of recording conditions for the expressive data. In the neutral partition of these settings only phonetically rich sentences are present, which represents 30% of the neutral speech contained in DC1 (approximately 2 hours of speech per speaker). The main motivations to reduce the amount of neutral data were: (1) to be able to train a good NTTS model with less data (since the selected sentences guarantee a good phonetic coverage) and (2) to assure a better balance between the neutral data and the expressive data.

For each selection of the expressive data, a different dataset configuration was designed. Therefore, DC2 contains only expressive data with high reverberation; DC3 contains expressive data recorded in studio quality; and finally, DC4 contains expressive data collected in both recording conditions. A detailed view of data configurations and audio samples are presented on our demo webpage.

3.2. Proposed Approaches

Our NTTS architecture is based on Tacotron2 [2] from Mozilla implementation. Tacotron2 is a state-of-the-art NTTS model that maps grapheme or phoneme sequences into mel spectrograms. The predicted mel spectrograms are synthesized using Griffin-Lim vocoder [16]. To accomplish the multi-speaker modeling, we added a speaker identity embedding layer. The speaker embedding output is then broadcast-concatenated to the decoder input.

To generate LRP space, we augmented the Tacotron2 architecture with different style encoders. The first architecture was based on a GST module with six tokens and four heads, and we name this as simple GST-Tacotron2 architecture. The choice for six tokens and four heads was an attempt to avoid the high degree of freedom that GST has. In the second one, we removed GST module and used only the Reference Encoder (RE-Tacotron2) without the last fully connected layer, similarly to [14]. Finally, the third architecture is based on Reference Encoder with an additional layer that receives as input the LRP space, generated by RE and classifies the expressivity of the reference audio in a supervised manner. In order to accomplish the style-guided modeling, we added Cross-Entropy Loss to the Tacotron2 loss function that measures the error of the classifier layer. We call this last one Style Guided RE-Tacotron2 (SGRE-Tacotron2). For all architectures, the style encoder layer’s output is broadcast-concatenated with the decoder input, similarly to the speaker embedding.

In order to generate controlled expressive speech using the LRP space, we used the centroid’s of each expressivity and, additionally, the expressive point with maximum distance in comparison with the neutral centroid. This last point is an attempt to generate an utterance “as expressive as possible” at inference
time; we refer to this point as Expressive Maximum Distance Point (EMDP).

3.3. Model Evaluation

A fundamental problem in the development of expressive speech synthesis models is how to evaluate the contribution of parameter or architectural changes to the final naturalness of speech. Typically, subjective perceptual evaluation is needed to assess the realism of synthesized speeches as well as the capability of the model in generating consistent speech styles. However, subjective perceptual evaluation is a laborious and time-consuming task, and it is not affordable for intermediate stages of development.

As an alternative to the subjective evaluation, we propose two objective metrics to evaluate how well our model fitted our target expressive style.

The ROC-AUC metric of a Logistic Regression trained on the learned LPR space can measure how well a linear model can classify the space. It is a proxy variable to show how separable are the expressive styles (neutral and expressive) in the LPR space. However, not only prosodic information is observed by the model. Since we have no control on what exactly is being modeled, it is not possible to state that a well separable space represents a model learned to distinguish only different prosodic styles.

As a second objective evaluation approach, we trained a robust classifier based on the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) features extracted from only Studio Quality data (DC3) and use this model to classify whether synthesized audio is expressive or not based only on acoustic features [9]. This set of features was elaborated aiming at a set of acoustic parameters that shows a good performance in affective computing tasks applied to speech. The GeMAPS features are characterized by: (1) their potential to index affective physiological changes in voice production, (2) their proven value in former studies as well as their automatic extractability, and (3) their theoretical significance in affect theory. The minimalistic set of features consists of 62 parameters extracted from 18 Low-level descriptors (LLD) based on frequency (pitch, jitter, formant’s frequency), energy/amplitude (shimmer, loudness, harmonics-to-noise ratio), and spectral (alpha ratio, hammarberg index, spectral slope, formant’s relative energies, and harmonic differences). Particularly, we used the extended version, called eGeMAPS, that has seven LLD added: spectral (MFCC and spectral flux) and frequency (Formant’s). As discussed in [9], cepstral parameters have proven to be highly successful in modeling affective states. In total, the extended set of features, eGeMAPS, contains 88 parameters. In our experiments, we trained a 5-fold Random Forest classifier that recognizes whether an audio is expressive or not with an accuracy of 98%, based on eGeMAPS features, the eGeMAPS classifier.

To assess the ability of our NTTS model to generate consistent expressiveness, we select 150 utterances, never seen in training, and classify the synthesized speech samples using the classifier. The synthesized speech samples are conditioned to the EMDP point, the eGeMAPS features are extracted from the output audio and then classified by the classifier. It is therefore expected that the classifier will be able to focus on prosodic features and indicate whether the synthesized speech is expressive or not.

4. Results

4.1. Low Dimensional Representations

Initially, we seek to evaluate different approaches in building the LRP space applied to the configuration of DC1 data. The GST-Tacotron2 architecture was the first to be evaluated, where some works have already used the approach to archetypal styles (such as Ekman’s emotions). We used the UMAP [17] technique to project the LRP space to a 2D representation and we analyze the distribution of styles in this lower space. Figure 1 shows the 2D-dimensional projection given by UMAP. The green points represent the representations of the expressive data and the blue points the neutral ones, both from Speaker 1, while the gray (Pool) points are the representations of the other speakers’ neutral data.

Although the architecture managed to concentrate the data labeled as expressive in a specific region of space, that same re-

---

### Table 1: Proprietary Dataset Available

<table>
<thead>
<tr>
<th>Speaker ID</th>
<th>Gender</th>
<th>Style</th>
<th>Recording Conditions</th>
<th>Data Partition</th>
<th>#Utterances</th>
<th>#Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker 1</td>
<td>Female</td>
<td>Expressive</td>
<td>Medium Quality</td>
<td>All Available</td>
<td>381</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Expressive</td>
<td>Studio Quality</td>
<td>All Available</td>
<td>265</td>
<td>0.29</td>
</tr>
<tr>
<td>Speaker 2</td>
<td>Female</td>
<td>Neutral</td>
<td>Studio Quality</td>
<td>All Available</td>
<td>7209</td>
<td>6.38</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>Neutral</td>
<td>Studio Quality</td>
<td>Phonetically rich</td>
<td>2389</td>
<td>2.24</td>
</tr>
<tr>
<td>Speaker 3</td>
<td>Male</td>
<td>Neutral</td>
<td>Studio Quality</td>
<td>All Available</td>
<td>17992</td>
<td>15.00</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>Neutral</td>
<td>Studio Quality</td>
<td>Phonetically rich</td>
<td>2543</td>
<td>2.61</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>Neutral</td>
<td>Studio Quality</td>
<td>All Available</td>
<td>5173</td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>Neutral</td>
<td>Studio Quality</td>
<td>Phonetically rich</td>
<td>1828</td>
<td>2.04</td>
</tr>
</tbody>
</table>
region still consisted of neutral data from all speakers. It was not possible to achieve an adequate modeling of the speech style in when listening to the artificial audios generated by conditioning the synthesis to several regions of the space (centroid’s or even the EMDP point), as well as the points described in the proposed approaches.

Similar to what was observed in [14], GST was not able to model different speech styles under the present conditions of experimentation. Unlike [13] for example, we don’t have a balanced dataset between different styles, and also we don’t use archetypal styles. Because of this, we have chosen to use only the Reference Encoder as our speech style encoder.

We note that, the RE-Tacotron2 architecture trained on DC1 was able to generate a separable space between the expressive and neutral data for all speakers, Figure 2. However, it is also noted that the generated LRP space is more sparse, having several agglomerations along with data of the same style. When listening to the generated audios, this architecture showed to be more effective to modify acoustic parameters in the synthesized speech by conditioning it to regions of neutral or expressive data. However, when conditioning to the EMDP point, the hearing quality of the synthesized speech seemed more associated with the recording condition of the expressive subset than with the speaking style itself.

Figure 2: UMAP projection of LRP space generated by RE-Tacotron trained on DC1.

We then performed the same experiment, using the SGRE-Tacotron2 architecture. We noticed that the generated LRP space guided in a supervised way is less separable than in the second experiment, but still more separable than in the first one, as illustrated by Figure 3.

In this experiment, we noticed that with the EMDP point, the artificial speech generated is closer to the expressive style while still modeling the different recording conditions. When using the expressive centroid for inference, there was no explicit modeling of the style. We, therefore, chose to follow the SGRE-Tacotron2 architecture in the subsequent experiments. Table 2 summarizes the conclusions of these three experiments.

4.2. Data diversity experiments

Based on the experiments in the LRP space described in the previous topic, we choose to continue using the SGRE-Tacotron2 architecture. Next, we tried to assess how different data conditions influence the technique’s ability to generate expressive artificial speech. First of all, we noticed that a large amount of data is not necessary if you have a smaller amount of phonetic richness data. In addition, a smaller amount of neutral data allows a better balance of the dataset in relation to expressive data. The architecture was then trained with the DC2 data. The generated LRP space is clearly separable, as shown in Figure 4. Apparently, a more balanced amount of expressive data helps the model to generate a more separable LRP space. In addition, as previously noted, the model is able to generate expressive speech by conditioning to the EMDP point but still jointly models the recording condition.

Figure 4: UMAP projection of LRP space generated by SGRE-Tacotron trained on DC2.

The LRP space generated by this approach allows a linear model to reach an 87.33% ROC-AUC in the validation set, showing itself to be a highly separable space. However, only 48% of the synthetic utterances conditioned on the EMDP point were classified as expressive by the eGeMAPS classifier, which
Table 2: Summary of Low Dimensional Representation experiments

<table>
<thead>
<tr>
<th>Model Architecture</th>
<th>Data Configuration</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>GST-Tacotron2</td>
<td>DC1</td>
<td>Not able control acoustic features with Tokens or generate a separable space</td>
</tr>
<tr>
<td>RE-Tacotron2</td>
<td>DC1</td>
<td>Can generate a separable space, but no certain of modeling prosodic information</td>
</tr>
<tr>
<td>SGRE-Tacotron2</td>
<td>DC1</td>
<td>Can guide LRP space to a better prosodic modeling but with no clear consistency in controlling expressivity</td>
</tr>
</tbody>
</table>

supports our hypothesis that the architecture models not only the style, but also the recording condition.

DC3 was used in our next test setup. In this case, both neutral and expressive speech have the same recording condition. This approach leads to a less separable LRP space, with a lower ROC-AUC value (81.74). Moreover, when listening to the synthetic utterances they sound less distinguishable with regard to the target style, Figure 5. In practice, this experiment resulted in audios close to the neutral style, even when conditioned to the EMDP point. Despite this, our eGeMAPS classifier model recognized 97.33% of the synthesized audios as expressive.

Figure 5: UMAP projection of LRP space generated by SGRE-Tacotron trained on DC3.

Our last experiment used all the available expressive data (DC4), maintaining the same SGRE-Tacotron2 architecture. The generated LRP space remains separable, but the studio quality expressive samples are clearly closer to the neutral samples than the medium quality ones, as shown in Figure 6. In this space, the linear model achieves a ROC-AUC of 85.52, and the eGeMAPS classifier model was able to identify 87.33% of the artificial utterances generated using the EMDP point. With this configuration, the artificial speech generated by the model sounds more expressive and with less recording condition, as if the recording condition highlighted the expressive data allowing the model to capture the prosody of expressiveness. Table 3 summarize all experiments reported in this topic.

Even though the first experiment resulted in a more separable space, it mostly models the recording condition itself. As a result, even when conditioning the model to the EMDP point, our eGeMAPS classifier could not recognize such synthesized speeches as actually expressive. On the other hand, when we train the model only on studio quality data, the generated space does not seem so separable, but the eGeMAPS model can identify the synthesized speeches as expressive. However, when listening to such audios, they are very similar to neutral ones, indicating a possible instability of the eGeMAPS classifier model. On the other hand, when we used all the expressive data in the third experiment, the SGRE-Tacotron2 was able to model less recording condition while still having a separable LRP space. It indicates that the variability of recording conditions for the same speech style can prevent only the recording condition from being modeled while still modeling prosody.

The results suggest that not even guided training can guarantee that prosodic features will be properly modeled. Finally, the presence of a second distinguishing factor among utterances, as in the first experiment, reinforces the architecture’s ability to model all speech aspects present in that group (prosodic characteristics and recording condition).

5. Conclusions and Discussions

The use of latent representations of audios (LRP) to model acoustic parameters of speech has been widely used to deal with expressive speech synthesis. Although this approach does manage to model speech characteristics, it did not explicitly guarantee the modeling of prosodic parameters in isolation under experimentation conditions of this work. Archetypal speech styles are associated with striking features in speech, allowing the model to separate its characteristics from one another easily. However, when dealing with styles typically used in real
life applications, the task becomes more challenging. Our experiments suggest that the model tends to model the audios’ most striking features, be them recording conditions or prosody. The variability of recording conditions in utterances of the same style, combined with a guided training of the LRP space, proved to be a promising approach to deal with styles that are not so easily distinguishable from each other.

Additionally, the present work is innovative in terms of expressive speech synthesis based on NTTS models applied to Brazilian Portuguese, and it can serve as a basis for future studies along the same line. Moreover, we have proposed the eGeMAPS Classifier as preliminary objective metric to evaluate expressive TTS models.

As a future work, we intend to continue studying the influence of different recording conditions on different architectures to help model real-life application speaking styles. We also intend to evaluate the use of prosodic variables to highlight such characteristics within the LRP spaces. Additionally, we intend to enrich our dataset with additional styles in order to perform similar experiments using two different style besides neutral one.

6. Acknowledgment

The authors would like to thank the Research and Development Institute CPQD and the Ministry of Science, Technology and Innovations for supporting and funding this project.

7. References


