Neural Speech Synthesis with Style Intensity Interpolation: A Perceptual Analysis

Noé Tits
Numediart Institute - UMONS
Mons, Belgium
noe.tits@umons.ac.be

Kevin El Haddad
Numediart Institute - UMONS
Mons, Belgium
kevin.elhaddad@umons.ac.be

Thierry Dutoit
Numediart Institute - UMONS
Mons, Belgium
thierry.dutoit@umons.ac.be

ABSTRACT

State of the art in speech synthesis considerably reduced the gap between synthetic and human speech on the perception level. However the impact of a speech style control on the perception is not well known. In this paper, we propose a method to analyze the impact of controlling the TTS system parameters on the perception of the generated sentence. This is done through a visualization and analysis of listening test results. For this, we train a speech synthesis system with different discrete categories of speech styles. Each style is encoded using a one-hot representation in the network. After training, we interpolate between the vectors representing each style. A perception test showed that despite being trained with only discrete categories of data, the network is capable of generating intermediate intensity levels between neutral and a given speech style.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Deep Learning; Speech Synthesis; Style Interpolation; Perception

ACM Reference Format:
Noé Tits, Kevin El Haddad, and Thierry Dutoit. 2020. Neural Speech Synthesis with Style Intensity Interpolation: a Perceptual Analysis. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20 Companion), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3371382.3378297

1 INTRODUCTION

In Human Agent Interaction, one of the main task is Text to Speech (TTS) allowing the agent to communicate information to the user. In recent years, deep learning based TTS systems allowed the generation of synthetic speech sounding very close to natural human speech. Currently research interests are focused on controlling para-linguistic dimensions such as the style of speech being generated [1, 5–7, 9].

Some systems able to learn latent representations of styles [2, 3, 8] have been successfully developed. These systems are able to interpolate in the space of this representation.

HRI '20 Companion, March 23–26, 2020, Cambridge, United Kingdom

ACM ISBN 978-1-4503-7057-8/20/03.
https://doi.org/10.1145/3371382.3378297

However, it is still not well known what are the effects of such control on the perception of the style by humans. The classical test to evaluate the synthesis subjectively being a Mean Opinion Score (MOS) test that only takes the naturalness into account. A MOS test is a subjective quality evaluation test in which subjects are asked to rate the quality of samples according to their opinion. The average of these is the MOS score.

In this paper, we study the capabilities of a deep learning based speech synthesis system to interpolate between different intensities of styles while being trained only on a dataset containing a neutral category and different categories with maximum intensity. We aim to study the perception of a controllable TTS system in terms of specific style categories measured continuously thanks to a score obtained by comparison of pairs of synthesized utterances.

2 DATASET

The dataset consists of recordings of a male english-speaking actor that was asked to read predefined sentences with different styles. This dataset was built to reproduce storytelling oriented styles of speech. The speech styles are: neutral, happy, sad/depressed, bad/mean, loud/from afar, whispered/from close, old.

For every style, recordings contain approximately 2 hours of speech. For more details, see the description in [8].

3 MODEL

The system consists of a multi-style TTS system with the possibility to control the intensity of style categories. It is a modified version of *Deep Convolutional Text-to-Speech* (DCTTS) [4], a deep learning based TTS system. In the modified version, it takes an encoding of the category at the input of the decoder. During training, a simple one-hot encoding is used, i.e. a code of 7 dimensions corresponding to the different styles. A 0 is assigned to all dimensions except for the style for which a 1 is assigned.

The text is represented with a sequence of N character embeddings of size e. Therefore, it is a matrix of shape (N, e). The 7-d vector is repeated N times to have a matrix of dimensions $(N, 7)$. These two matrices are concatenated to have a matrix of shape $(N, e + 7)$. And this matrix is fed to the rest of the architecture.

In other words, the style vector is *broadcast-concatenated* to the transcription embedding.

At synthesis stage, we can modify the intensity of a style category by interpolate between codes.

4 ANALYSIS OF THE IMPACT OF CONTROL VARIABLES ON STYLE PERCEPTION

In order to study if the DNN is able to interpolate in intensities without having seen intermediate levels of style encodings. In this

experiment, we interpolate between neutral and each style. Only the number corresponding to neutral and the category with which we interpolate are non-zero. The sum of the values at the positions corresponding to the neutral style and the target style is equal to 1.

Given two one-hot vectors v and w, and $\alpha \in [0, 1]$, the resulting encoding is

$$c = \alpha \cdot v + (1 - \alpha) \cdot w$$

To study the relationship between the control variables and the perception of styles in synthesized utterances, listening tests have been performed.

According to previous work, humans appear to be unreliable to assign an absolute value of intensity to a subjective concept. Humans are on the other hand better to compare and classify elements. It is related to the ordinal nature of emotion [10].

Inspired from this paper, we design our experiments based on the comparison of pairs of elements rather than asking participants to give absolute scores. We then analyze the correlation between the obtained scores and the interpolated parameters in order to study the latter's effect on the perception.

For each style, we synthesize 5 different utterances. We used the 5 first sentences of the standard Harvard sentences. It is a set of phonetically balanced sentences, i.e. the frequency of phonemes is the same as they appear in English. These are synthesized with 5 different intensity levels: 0, 0.25, 0.5, 0.75 and 1. This makes a total of 150 synthesized files. Our goal being to observe the effect of the interpolation for each style, we create 10 pairs of intensities per style for the comparison test.

For 6 styles, 5 sentences and 10 pairs of intensity levels, there is a total of 300 pairs.

The test is a multiple choice questionnaire in which the participants are presented the 300 pairs in a random order. They were then asked "Which sample sounds more *style*, A or B ?"

The possible responses were: ['A', 'B', 'Neither', 'They sound equally *style*']. The listening test was implemented using turkle[1], an open-source Mechanical Turk platform that can be run locally.

Each file will appear in 4 comparisons in one listening test. For each comparison, a score of 1 is assigned to the selected sentence and 0 to the other. If equal is selected, they both are attributed 0.5. If neither is selected, they both are attributed 0. For each file, the scores are then accumulated to obtain a score representing the *perceived intensity* of the style.

In the ideal case in which the control variable would exactly correspond to the perception of the participant, the score assigned to the intensities 0, 0.25, 0.5, 0.75 and 1, would be 0, 1, 2, 3 and 4.

5 RESULTS

To visualize the relationship between the perceived intensities and the control variables, we show a heatmap of the values of both variables by category and by subject in Figure 1. To quantify this relationship, we compute the Pearson Correlation coefficient between these two variables (Table 1).

The listening test was done on three participants. An expert in the field familiarized with the content of the dataset and the synthesis system (Subject 3) and two other people: one female (Subject 1) and one male (Subject 2).

[1]https://github.com/hltcoe/turkle

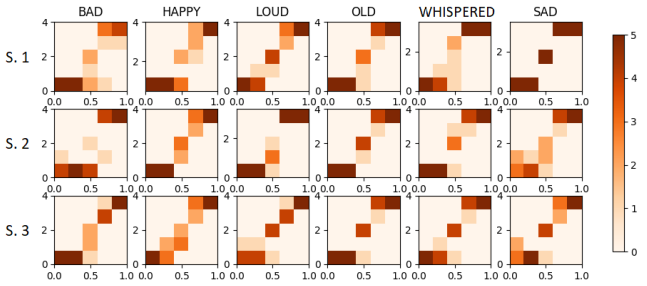

Figure 1: Heatmap of the number of associations between accumulated scores representing the perception of the intensity level of the style and control variables. From top to bottom, each line correspond to Subject 1, 2 and 3 respectively. The y-axis is the accumulated score from 0 to 4. The x-axis is the control variable. The color correspond to the number of times a participant associated x to y.

Table 1: Pearson Correlation Coefficient between control variables and perceived intensities by subject and by category

	bad	happy	loud	old	whisp.	sad	Overall
S. 1	0.862	0.914	0.953	0.931	0.913	0.947	0.913
S. 2	0.839	0.942	0.917	0.948	0.924	0.881	0.904
S. 3	0.944	0.946	0.929	0.929	0.930	0.900	0.929

An estimation of the mean and standard deviation of the Pearson Correlation Coefficient between the intensities and a random choice can be computed. It serves here as a random baseline. We sample 150 elements of a uniform discrete distribution and compute the Pearson correlation coefficient. By repeating this n times with $n = 10000$, we have a distribution with a mean of 0 ± 0.005 and standard deviation of 0.081.

The results in Table 1 show that the perception of style is highly correlated with the control parameters of the DNN and that the DNN is able to generate intermediate styles even though it had only seen discrete styles during training.

6 CONCLUSIONS

This paper studied the ability of a DNN TTS model to interpolate intermediate intensities of six different styles of speech, despite having been trained only on neutral and maximally styled utterances. The results show that a listening test based on comparison of pairs of audio files synthesized with different intensities of style lead to a perception score highly correlated to the intensity used for synthesis. However intermediate levels seem difficult to distinguish compared to extreme ones.

ACKNOWLEDGMENTS

Noé Tits is funded through a FRIA grant (Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture, Belgium)

REFERENCES

[1] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The Emotional Voices Database: Towards Controlling the Emotion

Dimension in Voice Generation Systems. *arXiv preprint arXiv:1806.09514* (2018).

[2] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al. 2018. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *arXiv preprint arXiv:1806.04558* (2018).

[3] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous. 2018. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *arXiv preprint arXiv:1803.09047* (2018).

[4] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2017. Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. *arXiv preprint arXiv:1710.08969* (2017).

[5] Noé Tits. 2019. A Methodology for Controlling the Emotional Expressiveness in Synthetic Speech-a Deep Learning approach. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 1–5.

[6] Noé Tits, Kevin El Haddad, and Thierry Dutoit. 2019. The Theory behind Controllable Expressive Speech Synthesis: A Cross-Disciplinary Approach. In *Human-Computer Interaction*. IntechOpen. https://doi.org/10.5772/intechopen.89849

[7] Noé Tits, Kevin El Haddad, and Thierry Dutoit. 2020. Exploring Transfer Learning for Low Resource Emotional TTS. In *Intelligent Systems and Applications*, Yaxin Bi, Rahul Bhatia, and Supriya Kapoor (Eds.). Springer International Publishing, Cham, 52–60.

[8] Noé Tits, Fengna Wang, Kevin El Haddad, Vincent Pagel, and Thierry Dutoit. 2019. Visualization and Interpretation of Latent Spaces for Controlling Expressive Speech Synthesis through Audio Analysis. In *Proc. Interspeech 2019*. 4475–4479. https://doi.org/10.21437/Interspeech.2019-1426

[9] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *arXiv preprint arXiv:1803.09017* (2018).

[10] G. N. Yannakakis, R. Cowie, and C. Busso. 2017. The ordinal nature of emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, Vol. 00. 248–255. https://doi.org/10.1109/ACII.2017.8273608