INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Analysis by Synthesis: Using an expressive TTS Model as Feature Extractor for Paralinguistic Speech Classification

Dominik Schiller[1,*], *Silvan Mertes*[1,*], *Pol van Rijn*[2], *Elisabeth André*[1]

[1]Human-Centered Artificial Intelligence, Augsburg, Germany
[2]Max-Planck-Institute for Empirical Aesthetics, Frankfurt, Germany

dominik.schiller@informatik.uni-augsburg.de, silvan.mertes@informatik.uni-augsburg.de, pol.van-rijn@ae.mpg.de

Abstract

Modeling adequate features of speech prosody is one key factor to good performance in affective speech classification. However, the distinction between the prosody that is induced by 'how' something is said (i.e., affective prosody) and the prosody that is induced by 'what' is being said (i.e., linguistic prosody) is neglected in state-of-the-art feature extraction systems. This results in high variability of the calculated feature values for different sentences that are spoken with the same affective intent, which might negatively impact the performance of the classification. While this distinction between different prosody types is mostly neglected in affective speech recognition, it is explicitly modeled in expressive speech synthesis to create controlled prosodic variation. In this work, we use the expressive Text-To-Speech model Global Style Token Tacotron to extract features for a speech analysis task. We show that the learned prosodic representations outperform state-of-the-art feature extraction systems in the exemplary use case of Escalation Level Classification.

Index Terms: GST Tacotron, Speech Analysis, Escalation Detection

1. Introduction

Identifying paralinguistic communicative goals from speech has a long research tradition [1]. In order to perform this classification in a supervised manner, one needs labels describing the communicative intent and some features describing relevant characteristics of the speech signal. Over the years, many handcrafted acoustic features have been compiled in exhaustive feature sets [2, 3]. Sets of this kind have shown good performance when applied to various acoustic classification problems, like emotion recognition [4]. However, the feature engineering and selection is a rather tedious and time-consuming task and can only be performed by experts with specific domain knowledge. To overcome those shortcomings, recent approaches often learn prosodic embeddings in an end-to-end manner directly from the speech signal, which makes this manual labour obsolete. Such state-of-the-art approaches yield a good performance on a number of paralinguistic tasks [5, 6]. During training they are optimized for a specific learning task that usually differs from the final classification task.

However, these feature extractors, that often are based on Deep Neural Networks (DNN), do not distinguish between the prosody induced by *what* is being said and *how* it is being said. The sentence-specific prosody is usually referred to as 'linguistic prosody'. This means that sentences are spoken differently due to differences in word stress (e.g., conTENT vs. CONtent, [7]), sentence focus (e.g. two WHITE shirts vs. TWO white shirts [8]) and broad pragmatic categories (e.g., statement or question sentences), to name just a few. On the other end, prosody relevant to the communication of paralinguistic goals is often called 'affective prosody' [9]. Figure 1 contains four spectrograms of two words spoken to be perceived as angry and neutral. The spectrograms for the same word contain strong similarities, while the spectrograms for different words are more distinct. Nonetheless, the angry recordings have a falling pitch contour, whereas this is not the case for the neutral recordings. The example highlights sentence-specific prosody is not directly related to the communication of paralinguistic goals. Moreover, sentence-specific prosody should not be confused with text semantics: While speech can be transcribed to text, it is not possible to perform semantic analysis directly on the sentence-specific prosody. This example illustrates that including sentence-specific prosody in feature embeddings might lead to less-performative features, as it might add irrelevant information. We therefore argue that the separate assessment of affective prosody could help improving the automatic identification of a speaker's paralinguistic intents.

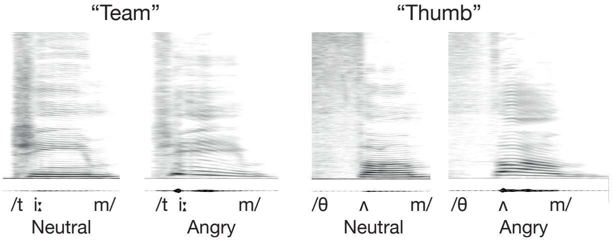


Figure 1: *Four recordings of two words for two emotions by the same speaker.*

While we do not know of any classification model for affective speech that takes the separation of different kinds of prosody into account, such a mechanism can be found in modern expressive Text-To-Speech (TTS) models. For example, the GST Tacotron [10] or Mellotron [11] models learn this distinction by allocating separate feature spaces for prosodic variation and text input during training. Thus, linguistic information is primarily excluded from the prosodic feature space in an implicit way, as this information is already included in the text embeddings.

In this work, we explore if prosodic embeddings of a trained GST Tacotron model can be used to improve the classification of a paralinguistic task. As a proof of concept, we take part in this year's Interspeech ComParE Escalation Sub-Challenge

[*] equal contribution

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-1587

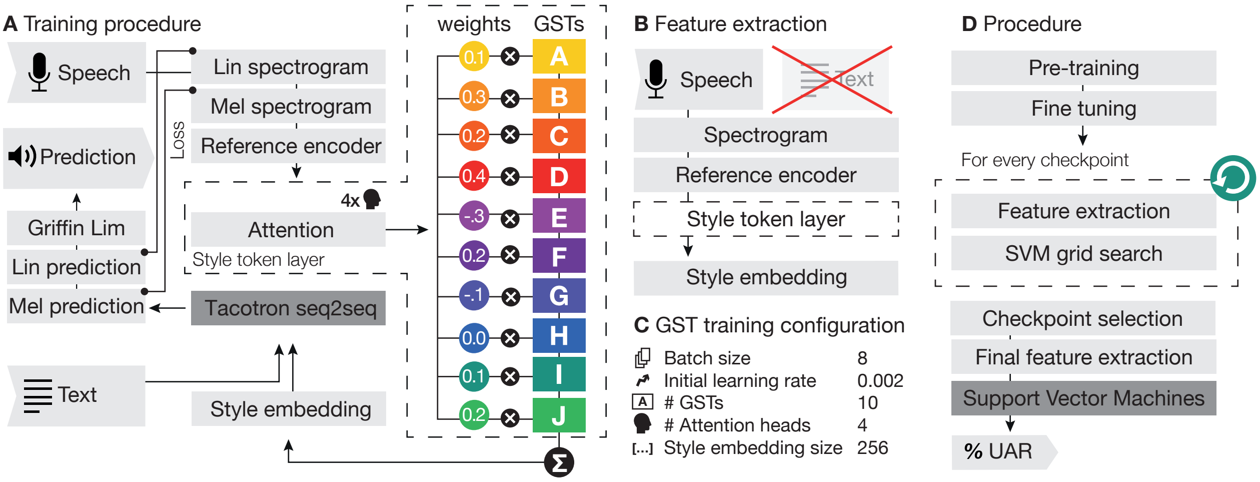


Figure 2: *(A) Simplified scheme of the GST Tacotron architecture. (B) Procedure of extracting the feature embeddings by the use of the trained GST Tacotron model. (C) Relevant configuration parameters of the GST Tacotron model for training. All other parameters were chosen as in the original implementation [10]. (D) Procedure of the model selection and configuration optimization during the experiments.*

[12], in which the level of escalation in Dutch dialogues has to be classified.

2. Related Work

Traditionally, large handcrafted feature sets like *GeMAPS* and *ComParE* [2, 3] have been used for classification tasks on the audio signal. More recently, deep learning approaches are applied to automatically learn feature embeddings from the raw signal. Prominent examples are *Deep Spectrum* and *AuDeep* [13, 14]. However, neither of these state-of-the-art approaches make a distinction between linguistic and affective prosody, because they do not take the transcript into account during training. On the other hand, purely text-based feature extraction approaches [15, 16, 17] only analyze the semantics of the transcribed text and not prosody.

Approaches that do make this distinction can be found in the field of TTS. While common TTS systems like Tacotron [18] only model the conversion of text to speech without directly taking the speaker's style or prosody into account, expressive TTS systems add a latent space to learn this prosodic variation. Thus, expressive TTS architectures are capable of modeling the prosodic style implicitly and learn more or less text-independent prosodic features.

GST Tacotron is one of the most well-known examples of such an expressive TTS system and builds upon the Tacotron framework, which is a sequence-to-sequence TTS model. The model has been shown to learn affective representations of speech [19]. To achieve the separation of the prosodic latent space, the model expands the original Tacotron architecture by a few additional components. First, a reference encoder is added to learn a compression function of the input mel spectrogram, resulting in a fixed-length prosodic representation of the speech audio. This so-called *Reference Embedding* of the input audio is then fed into an attention module and subsequently fed into a *Style Token Layer*. The attention module learns a mapping between the Reference Embedding and a bank of so-called *Global Style Tokens* (GSTs). Thus, the attention module assesses the contribution of each GST to the reference embedding. The weighted sum of the GSTs, called *Style Embedding*, along with the input text sequence, is then passed to a further encoder component, before being fed to Tacotron. A simplified scheme of the GST Tacotron architecture is shown in Figure 2A. During inference, the style embedding can either be extracted from existing reference speech samples or it can be directly controlled by modifying attention weights, which already has been applied to affective contexts in prior work. For a more in-depth description of GST Tacotron, we refer to the respective paper [10].

The architecture of GST Tacotron has been extended in other models, like Mellotron [11], that mainly focus on obtaining a more fine-grained control over prosody. For example, Mellotron is not only trained on recordings and transcripts, but also on extracted pitch contours from the respective recordings, giving the user full control over pitch. However, for our experiments, we use GST Tacotron because (i) it has the most minimal architecture, which speeds up the training process and (ii) the style embeddings hold all the information that is needed by the TTS system to add specific prosodic and style-specific characteristics to the output speech.

To the best of our knowledge, there exists no prior work that explores the use of TTS latent spaces for affective speech classification tasks.

3. Approach

We follow a two-step procedure. In a first stage, we train a GST Tacotron model on a TTS task. Our goal is, that the trained model is able to transform input text to speech that resembles the dataset. As described in the previous section, GST Tacotron implicitly learns a style embedding of the training data. Thus, after training, the reference encoder and style embedding layer can be used to extract style embeddings of audio data. While this style embedding was designed to be used to create speech with similar speaking styles, we use the embedding for classification, since it contains a more or less text-independent description of the speech prosody. In the second stage we extract the learned style embeddings and use them as a feature set for a paralinguistic classification task. This process is depicted in Figure 2B.

3.1. Pretraining and Fine-tuning

As training datasets for speech analysis tasks are typically sparse, we decided to add a pretraining stage to the GST Tacotron model. As a big portion of the learning problem of GST Tacotron consists of learning a text embedding and the decoding procedure, pretraining on a large dataset might substantially improve the training performance. We first trained the GST Tacotron model on a larger dataset, before finetuning it on the dataset of interest (see Section 4.1).

4. Experiments

To evaluate if our approach can be used to extract useful feature embeddings for speech analysis tasks, we take part in this year's Interspeech ComParE Escalation Sub-Challenge. The following section presents the experiments that were conducted.

4.1. Task and Datasets

The goal of the Escalation Sub-Challenge is classifying the escalation level of Dutch conversations. The provided corpus for the challenge is composed of the *Dataset of Aggression in Trains* (TR) [20] and the *Stress at Service Desk Dataset* (SD) [21]. Both datasets contain recordings of unscripted conversations in Dutch, recorded in varying quality. While the TR dataset is annotated regarding different levels of aggression, the SD dataset is labeled with respect to stress levels. Thus, for the Escalation Sub-Challenge corpus, both dataset annotations were mapped to a 3-level escalation scale. The dataset consists of 911 speech samples, each accompanied by a transcription of the spoken utterances. The corpus is split into three partitions: *train* (293 samples), *dev* (118 samples), and *test* (501 samples). For a more detailed description of the dataset, we refer to the official baseline paper [12].

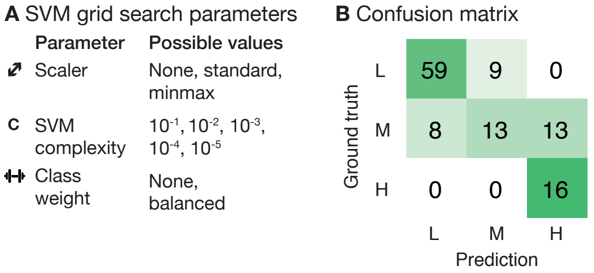
As described in Section 3.1, we included a pre-training stage to the GST Tacotron model, as the escalation corpus is comparably small for a TTS training task. Thus, we used the *Blizzard 2013* speech dataset for pretraining. The Blizzard 2013 dataset contains 9,741 segmented English utterance recordings spoken by the professional speaker Catherine Byers. We chose this dataset for pretraining for the following reasons: (i) GST Tacotron has shown to achieve promising results when trained with that dataset [10], (ii) the corpus contains prosodically varied speech, so the model would already learn lots of prosodic variation during pre-training and (iii) we used an English corpus as a Dutch speech corpus of this quality and size does not exist yet.

4.2. Training

The following section describes how we trained the GST Tacotron architecture and used the trained model to extract style embeddings of speech samples. The style embeddings were then used to fit a classifier to solve the task at hand.

4.2.1. GST Tacotron Training

The GST Tacotron model[1] was pretrained on the Blizzard 2013 dataset. After 380,000 epochs the pretraining was finished, as no further improvement could be observed. Subsequently, the model was fine-tuned on the escalation speech dataset. It should be noted, that for the training of the GST Tacotron model, the escalation labels are not used at all, as only a TTS task is learned. Thus, the speech data itself serves as ground truth, whereas the text transcriptions are given as input data. As *test* partitions are commonly used to represent data that is not available during training, it seems the most stringent to not use that partition for the training of the GST Tacotron model at all, although label data is not used at that stage. However, in practice there are many use cases in which a classifier can continue training as new, unlabeled data becomes available (e.g., in deferred, offline applications). Examples of such use cases include *Cooperative Machine Learning*, a paradigm, in which parts of a training dataset are labeled by humans and the remaining part of the dataset is annotated by a machine learning algorithm that was pre-trained on the human labels. In such a context, it would be unproblematic to incorporate the input data of the unlabeled test dataset into the GST Tacotron training process, given the fact that no escalation labels are used during that stage.


A SVM grid search parameters
B Confusion matrix
Ground truth
Prediction

Parameter	Possible values
Scaler	None, standard, minmax
C SVM complexity	10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}
Class weight	None, balanced

	L	M	H
L	59	9	0
M	8	13	13
H	0	0	16

Figure 3: *(A) SVM grid search parameters. (B) Confusion matrix on dev set, using style embeddings and the model that performed best on the dev set.*

Here, we address both scenarios in two different experiments. First, we performed the fine-tuning stage on the *train* and *dev* partitions only, second, we used all three partitions of the dataset (*train*, *dev* and *test*). The fine-tuning procedures were ended after further 340,000 epochs each, as again no further improvement could be observed. Relevant details of the training configuration of the GST Tacotron model can be taken from Figure 2C.

4.2.2. Classifier Training

After finishing the training of the GST Tacotron model, we discarded all layers that succeeded the style embedding layer. Thus, we obtain a network that outputs a style embedding for a speech input. We extracted the style embeddings for each sample of the escalation speech dataset. Since the aim of this work is to explore the use of such embeddings for a classification task, we trained a SVM on the extracted style embeddings of the GST Tacotron model to predict the escalation level. Unlike more sophisticated analysis models, such as DNN-based approaches, SVMs only allow limited room for fine-tuning and optimization, which makes it easier to obtain comparable results. We followed a similar SVM-based approach as in the baseline paper of the Escalation Sub-Challenge. In order to find the best fitting SVM configuration, we used a grid search approach, covering the most crucial configuration parameters. See Figure 2D for an overview of the approach. The parameters that were adjusted by the grid search are shown in Figure 3A.

[1] https://github.com/syang1993/gst-tacotron

Table 1: *Results of our approach on the Escalation Speech Dataset. The numbers printed in bold are the best values for the test and dev partitions respectively. The results on dev were achieved with a SVM trained on the train partition, the results on test were achieved with a SVM trained on train + dev*

		GSTs trained on *train* + *dev*	GSTs trained on *train* + *dev* + *test*
Selected with 5-Fold	Average Split UAR	71.5	71.4
	UAR on *dev*	68.9	67.6
	UAR on *test*	56.2	**61.1**
Selected by Performance on *dev*	UAR on *dev*	**75.0**	73.3
	UAR on *test*	58.6	59.2

4.3. Model Selection

As the GST Tacotron model was fine-tuned on the escalation speech dataset which holds a comparably low number of training samples, the model can easily overfit. Thus, we periodically saved checkpoints of the GST Tacotron model, and performed the SVM grid search for every single checkpoint separately. For the final model, we evaluated two different approaches for model selection. First, we trained the SVM for every checkpoint and grid search configuration on the *train* set only, and selected the model that performed best on the *dev* set in order to investigate how our approach achieves the best performance on *dev*. Second, we performed a 5-fold cross validation on the combination of *train* and *dev* partitions. Here, we selected the model with the best average performance on all five splits. In both cases, the best model was then trained on the *train* and *dev* sets combined to evaluate it on the *test* set. All the experiments were done twice, once with a GST Tacotron model trained on the *train* and *dev* partition and once on a GST Tacotron model trained on the whole dataset. The confusion matrix for the best *dev* model (trained on the *train* partition only) is shown in Figure 3B. It can be observed, that the trained SVM works best for high and low escalation samples, while mid level escalation speech is missclassified more frequently.

5. Results

The results of our experiments are shown in Table 1. As evaluation metric, we chose the *Unweighted Average Recall* (UAR), as this is the official metric for this task within the Interspeech ComParE Challenge. From the table, we can read that the configuration that performed best on the *dev* set reached a UAR of 75.0%. The best result on the *test* set was achieved with the embeddings that were produced by the GST Tacotron trained on the whole escalation speech dataset and led to a UAR of 61.1%. For both cases, the grid search estimated that using class balancing and a SVM complexity of 0.001 leads to the best results. However, the model that performed best on the *dev* set worked best when using a *Min-Max Scaler* for normalizing the training data, while the model that performed best on the *test* set used a *Standard Scaler*.

By using our approach, we outperform all state-of-the-art models that were reported in the baseline paper of the Escalation Sub-Challenge. There, the highest UAR on the *test* partition was 59.8% using *OpenXBOW*, a Bag-of-Audio-Words approach. All other feature extraction approaches that were reported in the baseline paper, namely *ComParE*, *OpenXBOW*, *Deep Spectrum*, *DiFE* and *End2You* reached even lower UARs.

6. Discussion

Our experiments showed that the features extracted from the pretrained GST Tacotron model can be effectively applied for the task of automatically detecting a speakers escalation level in natural speech. Moreover our approach managed to outperform all reported baseline approaches [12], which further substantiates its validity - especially when considering the fact that the results reported in the baseline paper, contrary to ours, are already selected with respect to their best performance on the *test* set. However, as can be seen in 1 the results on the respective sets may vary depending on the chosen methodology to determine the best model for the respective set. We can observe that adding the *test* set during the training of the GST embeddings has a positive impact on the results of the classification system on the *test* set. We argue that this implies that the GST Tacotron model actually learns dataset specific features that are relevant for affective recognition tasks on the respective data set, which suggests that the analysis and synthesis of human speech do not necessarily have to be seen as separate, distinct paradigms.

7. Conclusion & Outlook

In this work, we used GST Tacotron to learn feature embeddings for a speech analysis tasks. In the chosen evaluation use case, our approach outperformed all reported common state-of-the-art feature embedding methods and handcrafted feature sets. This demonstrates that expressive TTS systems offer a rich prosodic latent space for paralinguistic tasks and the GST Tacotron is a promising tool not only for speech synthesis, but also for speech analysis. Future research has to investigate how feature embeddings generated by our approach perform on different analysis tasks. We plan to apply the introduced approach to other datasets and use cases in the future.

8. Acknowledgements

This work has partially been funded by the European Union Horizon 2020 research and innovation programme, grant agreement 856879 (PRESENT).

9. References

[1] G. Fairbanks and W. Pronovost, “Vocal pitch during simulated emotion.” *Science*, vol. 88, no. 2286, pp. 382–383, Oct. 1938. [Online]. Available: https://doi.org/10.1126/science.88.2286.382

[2] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[3] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, “On the acoustics of emotion in audio: what speech, mu-

sic, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[4] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2013.

[5] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 147–151. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1238

[6] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[7] T. Gay, "Physiological and acoustic correlates of perceived stress," *Language and Speech*, vol. 21, no. 4, pp. 347–353, Oct. 1978. [Online]. Available: https://doi.org/10.1177/002383097802100409

[8] D. R. Ladd and R. Morton, "The perception of intonational emphasis: continuous or categorical?" *Journal of Phonetics*, vol. 25, no. 3, pp. 313–342, Jul. 1997. [Online]. Available: https://doi.org/10.1006/jpho.1997.0046

[9] M. Belyk and S. Brown, "Perception of affective and linguistic prosody: an ALE meta-analysis of neuroimaging studies," *Social Cognitive and Affective Neuroscience*, vol. 9, no. 9, pp. 1395–1403, Sep. 2013. [Online]. Available: https://doi.org/10.1093/scan/nst124

[10] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[11] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6189–6193.

[12] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen *et al.*, "The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates," *arXiv preprint arXiv:2102.13468*, 2021.

[13] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," 2017.

[14] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2017.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[16] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[17] R. A. Sinoara, J. Camacho-Collados, R. G. Rossi, R. Navigli, and S. O. Rezende, "Knowledge-enhanced document embeddings for text classification," *Knowledge-Based Systems*, vol. 163, pp. 955–971, 2019.

[18] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[19] P. van Rijn, S. Mertes, D. Schiller, P. Harrison, P. Larrouy-Maestri, E. André, and N. Jacoby, "Exploring emotional prototypes in a high dimensional tts latent space," *arXiv preprint arXiv:2105.01891*, 2021.

[20] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, "Automatic audio-visual fusion for aggression detection using meta-information," in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*. IEEE, 2012, pp. 19–24.

[21] ——, "An audio-visual dataset of human–human interactions in stressful situations," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 29–41, 2014.