ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097118

FINE-GRAINED EMOTIONAL CONTROL OF TEXT-TO-SPEECH: LEARNING TO RANK INTER- AND INTRA-CLASS EMOTION INTENSITIES

Shijun Wang[1]*, Jón Guðnason*[2]*, Damian Borth*[1]

[1]University of St.Gallen, Switzerland
[2]Reykjavik University, Iceland

ABSTRACT

State-of-the-art Text-To-Speech (TTS) models are capable of producing high-quality speech. The generated speech, however, is usually neutral in emotional expression, whereas very often one would want fine-grained emotional control of words or phonemes. Although still challenging, the first TTS models have been recently proposed that are able to control voice by manually assigning emotion intensity. Unfortunately, due to the neglect of intra-class distance, the intensity differences are often unrecognizable. In this paper, we propose a fine-grained controllable emotional TTS, that considers both inter- and intra-class distances and be able to synthesize speech with recognizable intensity difference. Our subjective and objective experiments demonstrate that our model exceeds two state-of-the-art controllable TTS models for controllability, emotion expressiveness and naturalness.

Index Terms— emotional TTS, emotion intensity control, speech emotion analysis

1. INTRODUCTION

Recent end-to-end Text-To-Speech (TTS) models [1, 2, 3, 4, 5] have the capacity to synthesize high-quality speech with neutral emotion. These models are, however, limited when it comes to expressing paralinguistic information such as emotion. It is critical to address this issue because expressing emotion in speech is crucial in many applications such as audiobook generation or digital assistants. Moreover, an additional challenge of current TTS models is the lack of fine-grained controllability of emotion on words or phonemes. Such a drawback results in inflexible control of speech, and failure to meet the context or users' intentions.

One straightforward strategy to express different emotions is by conditioning global emotion labels [6, 7]. However, synthesized speech from these models has monotonous emotional expression due to the condition of one global emotion representation. To achieve diverse emotion expression, models like GST [8] apply a token (a single vector) to represent the emotional style of a reference speech, then use this token to influence the synthesis. RFTacotron [9] is an extended work of GST. It uses a sequence of vectors instead of a single token to represent the emotion, which allows the improvement of the robustness and prosody control. Nevertheless, the nuance of references might be difficult to be captured by these models (e.g. one sad and one depressed reference might produce the same synthesized speech), due to a mismatch between the content or speaker of the reference and synthesized speech, which implies the inflexible controllability of these models.

A better approach to achieve fine-grained controllable emotional TTS is by manually assigning intensity labels (such as strong or weak happiness) on words or phonemes, which provides a flexible and efficient way to control the emotion expression, even for subtle variations. In [10, 11, 12, 13], Rank algorithms are used to extract emotion intensity information, by following the assumptions: i) speech samples from the same emotion class have similar ranks, and ii) intensity of neutral emotion is the weakest and all other emotions are ranked higher than neutral. Despite the production of recognizable speech samples with different emotion intensity levels, intra-class distance is neglected in these models. Specifically, during the training, samples belonging to the same emotion class (for instance, the strongest and weakest happiness) are arbitrarily considered the same. In practice, confusion could happen when we compare a median-level intensity speech with a strong- or weak-level one.

In this paper, we propose a TTS model, which outperforms the state-of-the-art fine-grained controllable emotional TTS models. The model is based on a novel Rank model, which is simple yet efficient for extracting emotion intensity information, by taking into account both inter- and intra-distance. Instead of performing rank on a non-neutral and a neutral sample, we use two samples augmented by Mixup [14]. Each augmented sample is a mixture from the same non-neutral and neutral speech. By applying different weights to non-neutral and neutral speech, one mixture contains more non-neutral components than the other one. In other words, one mixture's non-neutral intensity is stronger than that of the other. By learning to rank these two mixtures, our Rank model not only needs to determine the emotion class (inter-class distance), but also has to capture the amount of non-neutral emotion present in a mixed speech, i.e. intensity of non-neutral emotion (intra-class distance).

We summarize our contributions as: 1) we propose a fine-grained controllable emotional TTS model based on a

Authorized licensed use limited to: University of St. Gallen. Downloaded on November 23,2023 at 15:38:18 UTC from IEEE Xplore. Restrictions apply.

novel Rank model. 2) The proposed Rank model is simple and efficient to extract intensity information. 3) Our experimental results demonstrate that our TTS model outperforms two state-of-the-art fine-grained controllable emotional TTS models. Demo page can be found at `https://wshijun1991.github.io/ICASSP2023_DEMO/`.

2. APPROACH

We train two models. One is a Rank model that aims to extract emotion intensity representations. The other is a backbone TTS model used to generate speech.

2.1. Rank Model

Our Rank model is shown in Fig. 1. It maps the speech into intensity representations, then outputs a rank score regarding the emotion intensity. Input $\mathbf{X}$ is a concatenation of Mel-Spectrogram, pitch contour, and energy. $\mathbf{X}_{neu}$ indicates an input from neutral class, while $\mathbf{X}_{emo}$ represents an input from other non-neutral emotion classes. We then perform Mixup augmentation on the pair ($\mathbf{X}_{neu}$, $\mathbf{X}_{emo}$):

$$\begin{aligned} \mathbf{X}^i_{mix} &= \lambda_i \mathbf{X}_{emo} + (1-\lambda_i)\mathbf{X}_{neu}, \\ \mathbf{X}^j_{mix} &= \lambda_j \mathbf{X}_{emo} + (1-\lambda_j)\mathbf{X}_{neu}, \end{aligned} \tag{1}$$

where λ_i and λ_j are from Beta distribution $Beta(1,1)$.

The Intensity Extractor is then used to extract intensity representations. It first applies the same Feed-Forward Transformer (FFT) in [4] to process the input. We further add an emotion embedding to the output of FFT to produce intensity representations $\boldsymbol{I}^i_{mix}$ and $\boldsymbol{I}^j_{mix}$. This embedding is from a look-up table and depends on the emotion class of $\mathbf{X}_{emo}$. The addition of emotion embedding is to provide information on emotion class, because intensity might vary differently in various emotion classes.

From the intensity representations $\boldsymbol{I}^i_{mix}$ and $\boldsymbol{I}^j_{mix}$, we then average these two sequences into two vectors h^i_{mix} and h^j_{mix}. The original Mixup loss is further applied on them:

$$\begin{aligned} \mathcal{L}_{mixup} &= \mathcal{L}_i + \mathcal{L}_j, \text{where} \\ \mathcal{L}_i &= \lambda_i \text{CE}(h^i_{mix}, y_{emo}) + (1-\lambda_i)\text{CE}(h^i_{mix}, y_{neu}), \\ \mathcal{L}_j &= \lambda_j \text{CE}(h^j_{mix}, y_{emo}) + (1-\lambda_j)\text{CE}(h^j_{mix}, y_{neu}), \end{aligned} \tag{2}$$

and CE$(\cdot,\cdot)$ represents Cross Entropy loss, y_{emo} indicates labels for non-neutral emotion, while y_{neu} indicates neutral.

Despite the fact that Mixup has been demonstrated as an effective regularization method, there is little evidence showing it is sensitive to the intra-class distance. Thus, apart from $\mathcal{L}_{mixup}$ (inter-class), we need to introduce another loss to capture intra-class information. Inspired by [15], we first use a Projector (linear layers) to map the pair (h^i_{mix}, h^j_{mix}) to a scalar pair (r^i_{mix}, r^j_{mix}), where $r_{mix} \in \mathbb{R}^1$. r_{mix} is a score indicating the amount of non-neutral emotion present in speech, i.e. intensity. To force the model to correctly assign scores,

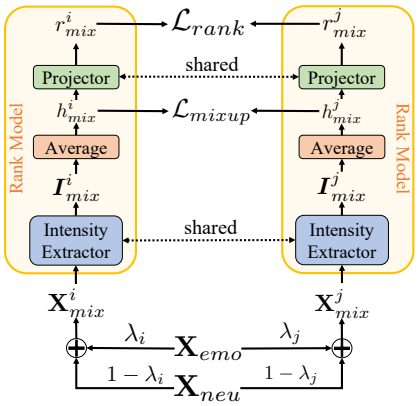


Fig. 1. The training of our Rank model. $\mathbf{X}_{emo}$ is a speech sample of non-neutral emotion classes, and $\mathbf{X}_{neu}$ is neutral. The Intensity Extractor produces intensity representations $\boldsymbol{I}^i_{mix}$ and $\boldsymbol{I}^j_{mix}$ given mixtures $\mathbf{X}^i_{mix}$ and $\mathbf{X}^j_{mix}$. h^i_{mix} and h^j_{mix} are two averaged vectors. $\mathcal{L}_{mixup}$ is a weighted cross entropy loss, while $\mathcal{L}_{rank}$ is to rank r^i_{mix} and r^j_{mix}, two scores, regarding the intensity of non-neutral emotion.

we first feed the score difference into a Sigmoid function:

$$p^{ij} = \frac{1}{1+e^{-(r^i_{mix} - r^j_{mix})}}, \tag{3}$$

then we apply the rank loss on it:

$$\mathcal{L}_{rank} = -\lambda_{diff}\log(p^{ij}) - (1-\lambda_{diff})\log(1-p^{ij}), \tag{4}$$

where λ_{diff} is a normalized result of $\lambda_i - \lambda_j$, which means if $\lambda_i > \lambda_j$, then $\lambda_{diff} \in (0.5, 1)$; if $\lambda_i < \lambda_j$, then $\lambda_{diff} \in (0, 0.5)$; if $\lambda_i = \lambda_j$, then $\lambda_{diff} = 0.5$.

As an example, if $\lambda_i > \lambda_j$ (non-neutral emotion presents more in $\mathbf{X}^i_{mix}$ compared to $\mathbf{X}^j_{mix}$), then $\lambda_{diff} > 0.5$, in this case, in order to decrease the rank loss in Eq. 4, the model needs to assign a bigger r^i_{mix} for $\mathbf{X}^i_{mix}$, to enable the Sigmoid output in Eq. 3 to be bigger than 0.5.

The intuition is forcing the model to correctly rank two samples that both contain non-neutral emotion. To achieve this, the intensity representation $\boldsymbol{I}_{mix}$ must convey information that can indicate the intensity of non-neutral emotion.

Lastly, we train our Rank model with the total loss:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{mixup} + \beta\mathcal{L}_{rank}, \tag{5}$$

where α and β are the loss weights.

2.2. TTS Model

We use FastSpeech2 [4] to convert the phonemes to speech, given intensity information. We maintain the original model configuration, except our Intensity Extractor is combined to provide intensity information.

The training of FastSpeech2 is shown in Fig. 2, and we only give a short description of each module here but refer the

Authorized licensed use limited to: University of St. Gallen. Downloaded on November 23,2023 at 15:38:18 UTC from IEEE Xplore. Restrictions apply.

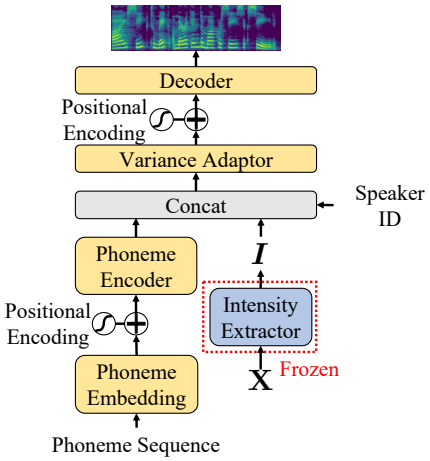


Fig. 2. The training of FastSpeech2, a trained Intensity Extractor is combined to provide intensity representations.

readers to the original paper [4] for more in-depth description. The Phoneme Encoder is to process phoneme and position information. The speaker ID is mapped to speaker embedding to represent speaker characteristics. The Variance Adaptor aims to predict pitch, energy and duration (frame length of each phoneme) based on the input. The decoder generates the final Mel-Spectrogram.

To incorporate intensity information, a pre-trained Intensity Extractor is frozen and integrated. And the Variance Adaptor uses phoneme, speaker information, and intensity representation $\boldsymbol{I}$ to predict pitch, energy, and duration. We set intensity representations for neutral emotion to zero, since we assume there is no intensity variation for neutral speech. One thing to point is that since the length of $\boldsymbol{I}$ is not equal to the phoneme length, we use Montreal Forced Aligner [16] to acquire intensity segments corresponding to each phoneme. Then the intensity segments are averaged to make the lengths of intensity representation and phoneme the same.

2.3. Training and Inference

Training: We first train our Rank model. Then the Intensity Extractor from the trained Rank model is frozen and combined during the training of FastSpeech2.

Inference: During inference, we expect to use phonemes and manual intensity labels to control the emotion intensity of synthesized speech. However, our Intensity Extractor can only output intensity representation from speech. In order to achieve controlling intensity with manual labels, we use the following strategy: with a trained Intensity Extractor, we first collect all intensity representations and their intensity scores. Then, we bucket all scores into several bins, where each bin denotes one intensity level (e.g. in our work, we use Min, Median and Max intensity levels, which means we apply three bins). After that, intensity representations corresponding to each intensity level are averaged into a single vector. Finally, we can map manual intensity labels to intensity representations during inference. This strategy is applied to each emotion class, therefore, we can find individual intensity representations by feeding emotion and manual intensity labels.

Implementation Details: We train the Rank model for 20k iterations with 1e-6 learning rate. For FastSpeech2, we use 250k iterations with 1e-4 learning rate. Adam optimizer is applied for both cases. In Eq. 5, α and β are respectively set as 0.1 and 1.0.

3. EXPERIMENTS

3.1. Experimental Setup

Dataset: EmoV-DB [17] is used as our dataset, it contains four speakers and five emotions (Amused, Angry, Disgusted, Neutral and Sleepy). The overall speech samples are around 7 hours and the sampling rate is 16KHz.

Data Preprocessing: Since FastSpeech2 is used as our backbone TTS, we need to feed Mel-Spectrogram, pitch and energy as inputs. We use 50-millisecond window and a 50 percent overlap ratio to extract energy and Mel-Spectrogram with 80 Mel Coefficients. PyWorld[1] is applied to extract pitch.

Baselines: We use FEC [11] and RFTacotron [9] as our baselines. Similar to our model, FEC has a Rank model and allows you to assign emotion intensity to each phoneme as well. To ensure a fair comparison, we replace the original Tacotron2 [2] in FEC with our FastSpeech2. RFTacotron transfers emotional style from a reference into the synthesized speech, which can be used to control the emotion intensity by applying reference samples with different intensities. We keep using the original Tacotron2 because its attention mechanism is the key part for emotion transfer.

Evaluation Setup: We use PWGAN [18] to convert generated Mel-Spectrograms to waveforms. To perform objective and subjective evaluations, for each emotion and speaker, we randomly select 5 unseen speech samples, by using their corresponding utterances, 90 utterances (speaker Josh has no data for Angry and Disgusted) are prepared for evaluation. 20 subjects participate in the subjective evaluations.

3.2. Emotion Intensity Controllability

In this section, we perform subjective evaluation to detect how recognizable of synthesized speech samples with different intensities (Min, Median or Max). Easily recognizable synthesized speech samples imply that we can efficiently control the emotion intensity by manually assigning intensity labels. Like in [19], for each utterance, we first synthesize three speech samples with three intensity levels, respectively. Then, three pairs (Min-Max, Min-Median and Median-Max) can be acquired and we ask subjects to select which one from a pair contains a stronger intensity. If the one that selected by subjects is the one synthesized with a stronger intensity, then it shows the emotion intensity can be appropriately controlled.

In FEC, the intensity rank scores are normalized in $(0, 1)$, thus, we refer to scores in $(0, 0.33]$ as Min, $(0.33, 0.66]$ as

[1] https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

Authorized licensed use limited to: University of St. Gallen. Downloaded on November 23,2023 at 15:38:18 UTC from IEEE Xplore. Restrictions apply.

Table 1. Rate Accuracy of Emotion Intensities. Min, Median and Max are three intensity levels. Subjects are asked to select the sample with the stronger intensity from a pair.

Emotion	Models	Intensity Pairs		
		Min -Median	Median -Max	Min -Max
Amused	RFTacotron	0.38	0.52	0.59
	FEC	0.63	0.58	0.63
	Ours	**0.66**	**0.60**	**0.74**
Angry	RFTacotron	0.53	0.59	0.60
	FEC	0.59	0.58	0.73
	Ours	**0.65**	**0.67**	**0.75**
Sleepy	RFTacotron	0.39	0.45	0.52
	FEC	0.56	0.54	0.64
	Ours	**0.65**	**0.73**	**0.83**
Disgusted	RFTacotron	0.48	0.51	0.53
	FEC	0.57	0.63	0.72
	Ours	**0.72**	**0.67**	**0.75**
Average	RFTacotron	0.45	0.52	0.56
	FEC	0.61	0.58	0.68
	Ours	**0.67**	**0.67**	**0.77**

Table 2. MCD and Naturalness MOS

Model	MCD (dB)	MOS
Ground Truth	/	3.9 ± 0.05
RFTacotron	5.21	3.49 ± 0.04
FEC	4.79	3.7 ± 0.04
Ours	**4.66**	**3.76 ± 0.03**

Median, and $(0.66, 1.0)$ as Max. Since RFTacotron is unable to assign intensity scores, we use our Rank model (Sec. 2.1) to find the strongest, median and the weakest samples from the dataset as references (based on the rank score r). We believe it is a fair comparison, because if our Rank model fails, then both RFTacotron and our model should fail.

As we can see from the results (Tab. 1), RFTacotron might not be efficient for performing intensity control, in some cases, the intensity difference is not easily perceivable. FEC improves a lot regarding the control of intensity, however, confusion happens when a median-level sample is in the pair. In other words, as we mentioned before, intra-class distance information might be partially lost in FEC. On the other side, our model performs the best compared with baseline models. It not only has the ability to synthesize Max- and Min-level samples, but also be capable of synthesizing recognizable Median-level speech samples.

3.3. Emotion Expressiveness

In this section, we conduct preference tests to evaluate whether models can express clear emotions. Since we don't consider intensity here, we only use synthesized samples with median-level intensity. For each utterance, we synthesize three median-level speech samples with our and two baseline models, respectively. Subjects are asked to select the one that conveys more clear emotion. If there is no detectable difference, they should choose "Same".

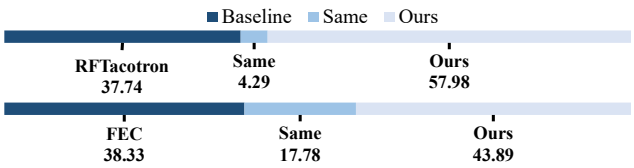


Fig. 3. Preference test for emotion expressiveness.

As we can see from the results (Fig. 3), our model significantly outperforms RFTacotron and subjects are barely confused, which suggests that our model's emotion expression is more clear than RFTacotron's. FEC also performs well on emotion expressiveness, but our model is preferred despite the same FastSpeech2 is used for both, which implies the benefit is caused by the intensity representation of our Rank model.

3.4. Quality and Naturalness Evaluation

We further evaluate quality and naturalness of synthesized speech samples. Objective measurement Mean Cepstral Distortion (MCD) [20], and subjective measurement Mean Opinion Score (MOS) are conducted for this evaluation.

Since we only focus on quality and naturalness here, and to be able to compare with ground truth speech, manual intensity labels are not used in this experiment. For FEC and our model, intensity representations are provided by their individual Rank models given ground truth speech. For RFTacotron, we use ground truth speech samples as references.

We report MCD and MOS results in Tab. 2. According to MCD results, both FEC and our model outperform RFTacotron largely, this might be because: 1) as opposed to transferring emotion from a reference, directly assigning intensity representations is easier for the model to generate good quality speech. 2) FastSpeech2 requires less data then Tacotron2 for a high quality result. Despite using the same FastSpeech2, the MCD of our model is slightly better than FEC. This is because our intensity representations might also bring benefits for high-quality synthesis. MOS scores (with 95% confidence intervals) reveal a similar phenomenon, where FEC and our model surpass RFTacotron greatly, while our model is slightly better than FEC.

4. CONCLUSION

In this paper, we propose a fine-grained controllable emotional TTS, based on a novel Rank model. The Rank model captures both inter- and intra-class distance information, and thus is able to produce meaningful intensity representations. We conduct subjective and objective tests to evaluate our model, the experimental results show that our model surpasses two state-of-the-art baselines in intensity controllability, emotion expressiveness and naturalness.

Authorized licensed use limited to: University of St. Gallen. Downloaded on November 23,2023 at 15:38:18 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] Jose M. R. Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron C. Courville, and Yoshua Bengio, "Char2wav: End-to-end speech synthesis," in *ICLR*, 2017.

[2] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018.

[3] Hyunseung Chung, Sang-Hoon Lee, and Seong-Whan Lee, "Reinforce-aligner: Reinforcement alignment search for robust end-to-end text-to-speech," in *Interspeech*, 2021.

[4] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *ArXiv*, vol. abs/2006.04558, 2021.

[5] Xu Tan, Tao Qin, Frank K. Soong, and Tie-Yan Liu, "A survey on neural speech synthesis," *ArXiv*, vol. abs/2106.15561, 2021.

[6] Younggun Lee, Azam Rabiee, and Soo-Young Lee, "Emotional end-to-end neural speech synthesizer," *ArXiv*, vol. abs/1711.05447, 2017.

[7] Paarth Neekhara, Shehzeen Samarah Hussain, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley, "Expressive neural voice cloning," *ArXiv*, vol. abs/2102.00151, 2021.

[8] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *ICML*, 2018.

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

[10] Xiaolian Zhu, Shan Yang, Geng Yang, and Lei Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 192–199, 2019.

[11] Yi Lei, Shan Yang, and Lei Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 423–430.

[12] Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[13] Bastian Schnell, "Controllability and interpretability in affective speech synthesis," Tech. Rep., EPFL, 2022.

[14] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *ArXiv*, vol. abs/1710.09412, 2018.

[15] Shijun Wang and Damian Borth, "Zero-shot voice conversion via self-supervised prosody representation learning," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 01–08.

[16] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *INTERSPEECH*, 2017.

[17] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *ArXiv*, vol. abs/1806.09514, 2018.

[18] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, 2020.

[19] Sarath Sivaprasad, Saiteja Kosgi, and Vineet Gandhi, "Emotional prosody control for speech generation," in *Interspeech*, 2021.

[20] Robert F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, pp. 125–128 vol.1, 1993.

Authorized licensed use limited to: University of St. Gallen. Downloaded on November 23,2023 at 15:38:18 UTC from IEEE Xplore. Restrictions apply.