Interspeech 2024
1-5 September 2024, Kos, Greece

Emotion Arithmetic: Emotional Speech Synthesis via Weight Space Interpolation

Pavan Kalyan, Preeti Rao, Preethi Jyothi, Pushpak Bhattacharyya

IIT Bombay

{1900201240@, prao@ee, pjyothi@cse, pb@cse}.iitb.ac.in

Abstract

While the idea of task arithmetic has been shown to be useful to steer the behaviour of neural models for NLP and vision tasks, it has not yet been used for speech. Moreover the tasks studied have been restricted to text classification and generation, and image classification. We extend the idea of task vectors to emotional speech synthesis in this work. We build emotion vectors by subtracting the weights of a pre-trained model from the weights of the same model after fine-tuning for a given emotion. These emotion vectors can be modified or combined through arithmetic operations such as negation and addition, with the hope of steering the behaviour of the resulting model accordingly in the generation of emotional speech. We also show that the emotion vector can achieve the desired transfer of emotion to a speaker not seen during training.

Index Terms: speech synthesis, emotions, task arithmetic

1. Introduction

Recent work has demonstrated that neural networks can be modified by interpolating between parameters of fine-tuned models, a technique known as *task arithmetic* [1]. Most studies focus on text and image modalities, but the effectiveness of task arithmetic for speech-based models is yet to be evaluated [2]. In this study, we show that weight-space interpolation can effectively modify the behaviour of text-to-speech systems in terms of generated speech emotion. We introduce *emotion vectors*, an extension of task vectors. These emotion vectors can be scaled to vary the intensity of the generated emotion, negated to express opposite emotions, combined to generate complex emotions, and transferred to speakers without emotional data.

While current text-to-speech (TTS) systems have achieved human-level naturalness, synthesizing emotionally expressive speech poses a challenging problem due to the complex spectrum of emotions. Each sentence can be spoken with one or multiple emotions by adjusting certain prosodic features of the speech [3]. The emotions we convey can vary in intensity, be discrete, or even a mixture at times. Consequently, obtaining annotated data for multiple emotions is a difficult and ambiguous task. This motivates the use of a predefined structured model of emotions, such as Plutchik's model [4]. Plutchik proposed eight primary emotions that combine to form complex emotions. These primary emotions can vary in intensity, resulting in a wide range of different emotions. We apply these principles to emotion vectors to generate the desired emotional speech.

Existing emotional TTS models use discrete emotion labels [5], textual prompts [6], or continuous emotion dimensions like valence (the pleasantness of a stimulus), arousal (the intensity of emotion provoked by a stimulus), and dominance (the degree of control exerted by a stimulus) [7]. Utilizing discrete emotion labels limits the model's ability to synthesize complex emotions. Acquiring textual prompts with emotional data is costly and often unavailable for most languages. Mapping between valence, arousal, and dominance values for complex emotions can be challenging and requires domain knowledge. Some previous studies rely on the structural model of emotions to synthesize emotional speech [8, 9]. These studies employ rank-based methods or emotion-embedding conditioning to adjust the intensity of emotions or generate speech with a mixture of emotions. More recently, emotionally separable prosody embeddings were proposed, along with manipulating these embeddings according to Plutchik's model to generate a wider range of emotions [10].

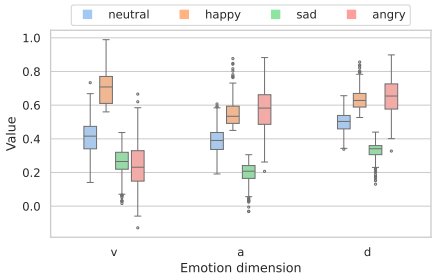

Figure 1: *Valence (v), arousal (a), and dominance (d) values of 300 speech samples from the ground truth neutral, happy, angry, and sad splits. This plot will serve as a reference for interpreting trends of v, a, and d values in generated speech.*

We present a novel approach to control the behaviour of speech-synthesis models to exhibit varying emotions. We employ the expressive TTS dataset known as Storynory [11] to train a variational autoencoder-based TTS system (VITS) [12]. This model is referred to as the pretrained model. We proceed to fine-tune the pretrained TTS model independently for each emotion, thereby creating a finetuned model for each emotion. By subtracting the weights of the pretrained model from the finetuned model, we obtain *emotion vectors*. These vectors are then utilized in various operations during inference time in order to generate emotional speech. Our contributions are:

1. Proposing emotion vectors as a simple, efficient, and effective method for controllable emotional speech synthesis[1].

[1]Code and samples: https://tinyurl.com/hnmz987n

10.21437/Interspeech.2024-2311

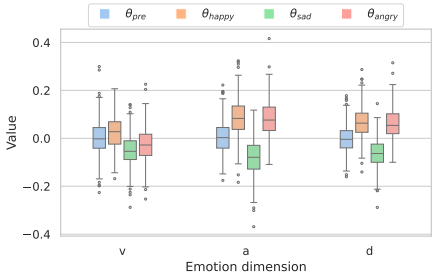

Figure 2: *Difference in v,a,d values between speech from fine-tuned models for each emotion and ground truth neutral speech samples. The distribution of these differences follows the same trend for each emotion as depicted in Figure 1.*

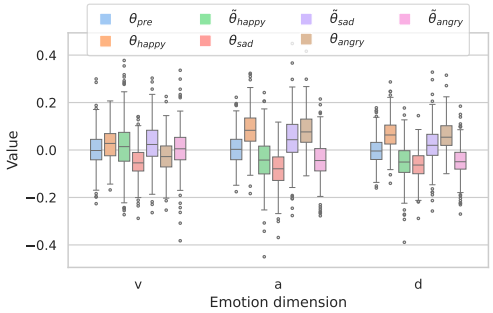

Figure 3: *Comparison of v, a, d and values of the speech generated by θ_e and the corresponding speech generated by $\tilde{\theta}_e$, w.r.t. ground truth samples. The distribution of difference values for θ_e exhibits a similar pattern to the emotion that is opposite to e.*

2. Application of emotion vectors to vary the intensity of emotions, negation, combination, and transferring emotions to an unseen speaker during inference.

2. Methodology

2.1. Dataset

All experiments used the Storynory dataset, a TTS dataset with 32 hours of expressive speech from children's storytelling. The TTS system was trained on the entire dataset to get a pretrained model. This pretrained model was then fine-tuned for each emotion. Emotion labels for the dataset were determined using a speech emotion classifier [13]. The classifier assigns a probability to each emotion (neutral, happy, anger, sad, surprise, fear, disgusted), with the highest probability emotion being the label for each speech segment. Only the emotions Happy, Sad, and Anger had more than 2 hours of speech after labelling. Therefore, speech samples (utterances) with these emotions were selected and filtered to retain only samples with high-confidence labels. For the Angry emotion, a threshold of 0.95 was chosen on the probability produced by the emotion classifier. Out of the 1569 angry samples, 1402 were used as the train set, 67 as the validation set, and 100 as the test set. The Happy and Sad emotions had around 10 hours of speech before filtering. These speech samples were initially passed through a three-dimensional emotion recognition model (VAD) [14], which predicted valence (v), arousal (a), and dominance (d) for each speech sample. Subsequent filtering based on these values for both Happy and Sad emotions resulted in 2 hours of high-quality speech. The train-validation-test split for Happy was 837, 33, and 100 samples respectively, and for Sad it was 1252, 105, and 100 samples, respectively. The speech emotion classifier categorized around 1 hour of data as Neutral. This data was further filtered using the probability score (> 0.95) and duration (> 3 sec). The neutral data is used for evaluation along with the 300 samples reserved for the test set and 69 for the validation set.

2.2. Emotion vectors

We utilize VITS as the underlying architecture to perform all our experiments. The model is trained on the Storynory dataset for 180K steps and is considered to be our pretrained model. We further fine-tune all parameters of the pretrained model for each emotion for an additional 25k steps to acquire the finetuned models. By subtracting the weights of the pretrained model from the fine-tuned model for a specific emotion, we obtain the emotion vector[2] for that emotion. Let $\theta_{\text{pre}} \in \mathbb{R}^d$ denote the weights of the pretrained model and $\theta_{\text{e}} \in \mathbb{R}^d$ be the weights of the model after fine-tuning on emotion e. The emotion vector $\phi_{\text{e}} \in \mathbb{R}^d$ is obtained by element-wise difference between θ_{e} and θ_{pre}, i.e., $\phi_{\text{e}} = \theta_{\text{e}} - \theta_{\text{pre}}$. These emotion vectors can be applied to any model parameters θ via element-wise addition, with an optional scaling factor α to get the resultant model as $\theta_{new} = \theta + \alpha\phi_{\text{e}}$. When $\alpha = 1$, the resultant model is one that is fine-tuned on that emotion. These emotion vectors are used to perform various arithmetic operations as detailed in the subsequent sections.

3. Experiments and Results

We conduct three experiments using the emotion vectors. The selection of α for each experiment is such that the Word Error Rate (WER) of the generated speech of sentences from the evaluation set w.r.t. the ground truth transcript is less than 0.3.

1. **Emotion negation.**
   Plutchik's model postulates that each primary emotion has an opposite counterpart. We exploit this insight to generate speech expressing sadness using the happiness emotion vector, and pleasant speech using the angry emotion vector. Let $\tilde{\theta}_e$ be the weights of the model for the opposite emotion of e, then $\tilde{\theta}_e = \theta_{pre} - \alpha\phi_{\text{e}}$. We chose α as 0.35 for all emotion vectors.

2. **Varying emotion intensity.**
   To vary the intensity of each emotion, we modify the scaling factor (α). Let θ_e^α denote the parameters of a model that generates speech in emotion e with intensity α, then $\theta_e^\alpha = \theta_{pre} + \alpha\phi_{\text{e}}$.

3. **Transfer to Unseen Speakers.**
   We add ϕ_{e} to the VITS model trained on the LJSpeech dataset [15] (neutral speech) to generate speech in emotion e. Let θ_{pre}^{ljs} denote VITS trained only on LJSpeech, and θ_e^{ljs} denote VITS (corresponding to LJSpeech speaker) that can generate speech in emotion e. Then, $\theta_e^{ljs} = \theta_{pre}^{ljs} + \alpha\phi_{\text{e}}$. α is 0.55 for angry, 0.4 for sad and 0.36 for happy emotion.

[2]Previous studies have referred to it as a vector due to its usefulness in performing mathematical operations to create new models.

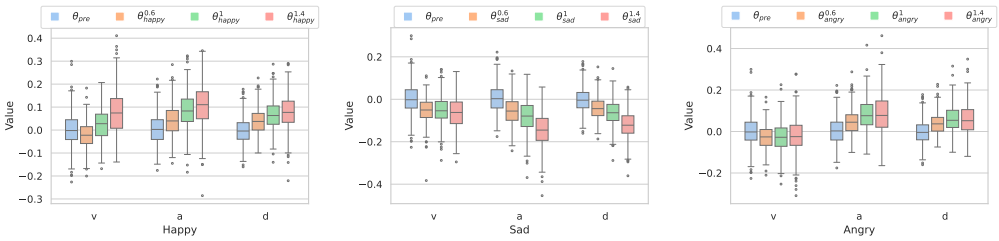

Figure 4: *Differences in v,a,d values w.r.t. ground truth when scaling the intensity of happy (left), sad (middle) and angry (right).*

3.1. Results

We conduct evaluations to assess 1) effectiveness of fine-tuned models in generating emotional speech 2) ability to generate the opposite emotion by negating the original emotion 3) impact of scaling α on the intensity of the emotion 4) capability of transferring emotion to an unseen speaker. In all of our evaluations, we utilize a neutral test set created in Section 2.1.

For objective evaluations, we feed the text into the desired model and produce corresponding speech samples. These speech samples are then assessed for emotion and intelligibility by passing them through the VAD model to predict the utterance v, a, d values and ASR transcripts. We compare the v, a, and d values of the model with those of the corresponding neutral ground truth samples, and plot the difference[3]. Additionally, we report the WER with reference to the ground truth text by using the Whisper-large v2 model [16] on the generated sample. The distribution of the difference in v, a, and d values with respect to the ground truth neutral speech samples is expected to adhere to the patterns displayed in Figure 1, which represents the distribution of v, a, and d values extracted from ground truth speech samples for each emotion. For subjective evaluations, we present each system to each evaluator at least five different text samples. We perform these evaluations with a total of 30 evaluators, resulting in 150 ratings per system.

1. **Effectiveness of fine-tuned models.**
   Figure 2 depicts the effectiveness of fine-tuned models in producing emotional speech. The plot shows increased valence, arousal, and dominance for θ_{happy} compared to θ_{pre}, while θ_{sad} exhibits a notable decrease in these dimensions. Additionally, θ_{angry} has lower valence and higher arousal and dominance than θ_{pre}. During the subjective evaluation, the assessors are presented with an audio sample and tasked with independently rating its anger, happiness, and sadness on a scale ranging from 1 to 5. A rating of 1 signifies not happy, whereas a rating of 5 indicates extreme happiness, for instance. Figure 5 demonstrates that the ratings consistently lean towards higher values on that emotion scale for which the model underwent fine-tuning.

2. **Ability to generate the opposite emotion.**
   Figure 3 shows that negating the emotion vector helps generate speech for the opposite emotion. The valence of $\tilde{\theta}_{happy}$ is lower than θ_{happy}. In contrast, $\tilde{\theta}_{sad}$ has higher valence than θ_{sad} and similar to θ_{happy}. This trend is also observed for anger. This pattern is consistent for arousal and dominance, showing the effectiveness of emotion vectors in synthesizing speech with opposite emotions. To evaluate this, we use the same subjective test as described above. In Figure 5, we compare θ_e and $\tilde{\theta}_e$. $\tilde{\theta}_{happy}$ receives higher ratings on sadness and anger scales compared to the happiness scale. $\tilde{\theta}_{angry}$ and $\tilde{\theta}_{sad}$ receive the highest ratings on the happiness scale.

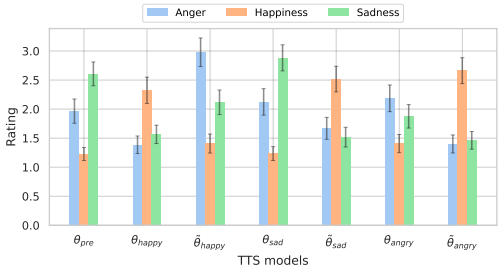

Figure 5: *Comparison of ratings by human evaluators for each emotion model on anger, happiness, and sadness scales. Ratings of models obtained via negation are also shown.*

3. **Impact of scaling alpha on the intensity of the emotion**
   Figure 4 presents the objective results of the experiment. The difference in v, a, d values compared to the ground truth increases as the intensity of emotion is scaled in the direction of the ground truth emotional samples. The v, a, d values increase as the happy emotion vector is scaled from $\alpha = 0.6$ to $\alpha = 1.4$. Similarly, these values decrease as the sad vector is scaled. For angry emotion, the valence decreases while the a, d value increases. These results are subjectively validated for sadness. During subjective evaluation, the evaluator rates four audios on a sadness scale of 1 (not sad) to 5 (extremely sad). Figure 6a shows that the sadness ratings for $\theta^{1.4}_{sad}$ are higher than θ^{1}_{sad} and $\theta^{0.6}_{sad}$.

4. **Transferability to Unseen Speakers**
   We performed only subjective evaluations here as the ground truth happy, sad, angry, and neutral samples are not available for the LJSpeech speaker and therefore, reference v,a,d scores are not available for an objective comparison. We present the evaluator with two audios, one generated from θ^{ljs}_{happy} and another from θ^{ljs}_{pre} and ask two questions 1) Which audio is happier? 2) Is the speaker in both audios the same or different? Figure 6a and b show the results for these subjective tests. Figure 6b shows that the resulting audios after adding the happy emotion vector are in fact happier, but Figure 6c

[3]All distributions are statistically different from each other (Mann-Whitney test, $p < 0.001$). Detailed statistics are provided on the sample page: `https://tinyurl.com/hnmz987n`

indicates that some of the speaker identity is getting lost in the process of this emotion transfer.

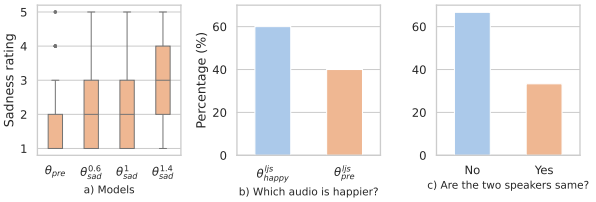

Figure 6: *a) Human ratings on the sadness of speeches generated by θ^{α}_{sad} as we scale the intensity via α. b) and c) show the preference of evaluators for two subjective questions asked when evaluating emotion transfer to an unseen speaker.*

It is important to report MOS for naturalness and WER as presented in Table 1. The results of both WER and MOS suggest that emotion vectors enable synthesis of emotional speech, while maintaining its overall naturalness and intelligibility.

Table 1: *MOS (95% confidence intervals) and WER*

Systems	MOS (CI)	Wer	Systems	MOS (CI)	Wer
GT	3.96(0.19)	0.17	θ_{pre}	3.48(0.21)	0.26
θ_{happy}	3.72(0.20)	0.16	$\tilde{\theta}_{happy}$	3.39(0.25)	0.33
θ_{sad}	3.62(0.21)	0.17	$\tilde{\theta}_{sad}$	3.67(0.20)	0.31
θ_{angry}	3.36(0.23)	0.18	$\tilde{\theta}_{angry}$	3.71(0.20)	0.26
θ^{ljs}_{pre}	3.68(0.19)	0.13	θ^{ljs}_{happy}	3.12(0.19)	0.15

4. Discussion

Combination of primary emotions is also a crucial aspect of the structural model of emotion. We have also attempted to combine emotions according to the structural model for three complex emotions: bittersweet (happy + sad), pride (happy + angry), and envy (sad + angry). The results of the objective evaluation of these combinations can be seen in Figure 7. Although the interpretation of the v, a, and d values for these complex emotions is not entirely clear, it is evident that these distributions differ from those of their primary emotions. Furthermore, the valence, which indicates the positivity or negativity of the emotion, follows the expected trend. For example, pride has a high valence, similar to happiness, while envy has a very low valence compared to sadness and anger. However, the subjective evaluations for these complex emotions is very ambiguous and often with low-agreement across evaluators [17].

All previous experiments involved fully fine-tuning the pretrained model for a specific emotion. We also experimented with fine-tuning the individual modules of the VITS architecture, such as the decoder, posterior encoder, prior encoder, and duration predictor. The prior encoder contains a text encoder and a normalizing flow module. Fine-tuning each module and freezing the others had no effect on the generated speech. However, when we fine-tuned only the text encoder and froze the rest of the modules, the model produced emotional speech. The text encoder takes phonemes as input and predicts a hidden representation h. This hidden representation is used in the normalizing flow to predict the latent representation z, which is further

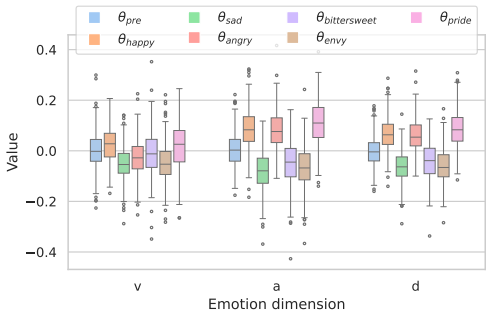

Figure 7: *Trends in difference of v,a,d values w.r.t ground truth neutral samples for combination of primary emotions*

used by the decoder. The hidden representation h is also used in the duration predictor to predict the duration, which is important for prosody modelling. The VITS text encoder is a Transformer block with 1d convolutional layers as feedforward layers. Interestingly, by solely fine-tuning these 1D convolutional layers of the text encoder while freezing all other modules, it is possible to generate emotional speech that is almost equivalent to speech obtained through full-finetuning of the model. Comparing θ_{happy} ($\approx 36M$ parameters) and θ^{ffn}_{happy} (model obtained by finetuning only the feedforward layers $\approx 5M$ parameters) in Figure 8, the v, a, d of these models for each emotion show very similar distributions, indicating that these are the truly impactful layers for emotion arithmetic.

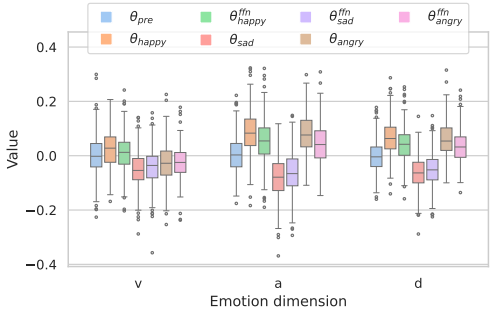

Figure 8: *Distribution of difference in v,a,d values for θ_e (full fine-tuning) vs θ^{ffn}_e (fine-tuning only feed-forward layers).*

5. Conclusion

We introduced a novel method to synthesize emotional speech by manipulating emotion vectors. This approach can replicate emotional speech with different (1) primary emotions, (2) polarity, (3) intensity levels, and (4) transferability to an unseen speaker. Through perceptual assessments, our method showcases a high degree of naturalness of speech along with the recognizability of the intended emotion. The arithmetic operations performed on emotion vectors involve solely the addition or subtraction of model weights, making them computationally efficient when compared to alternative techniques requiring additional fine-tuning. We leave the exploration of these emotion vectors for cross-lingual and cross-speaker transfer while retaining speaker characteristics as future work.

6. References

[1] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," *ArXiv*, vol. abs/2212.04089, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254408495

[2] H. Zheng, L. Shen, A. Q. Tang, Y. Luo, H. Hu, B. Du, and D. Tao, "Learn from model beyond fine-tuning: A survey," *ArXiv*, vol. abs/2310.08184, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263908791

[3] S. J. L. Mozziconacci, "Prosody and emotions," *Speech Prosody 2002*, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:14823425

[4] R. Plutchik, "A psychoevolutionary theory of emotions," *Social Science Information*, vol. 21, pp. 529 – 553, 1982. [Online]. Available: https://api.semanticscholar.org/CorpusID:144109550

[5] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," *ArXiv*, vol. abs/1711.05447, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:38104552

[6] Y. Leng, Z. Guo, K. Shen, X. Tan, Z. Ju, Y. Liu, Y. Liu, D. Yang, L. Zhang, K. Song, L. He, X.-Y. Li, S. Zhao, T. Qin, and J. Bian, "Prompttts 2: Describing and generating voices with text prompt," *ArXiv*, vol. abs/2309.02285, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261557296

[7] S. Kosgi, S. Sivaprasad, N. Pedanekar, A. K. Nelakanti, and V. Gandhi, "Empathic machines: Using intermediate features as levers to emulate emotions in text-to-speech systems," in *North American Chapter of the Association for Computational Linguistics*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249438639

[8] K. Zhou, B. Sisman, R. K. Rana, B.W.Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, vol. 14, pp. 3120–3134, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251492999

[9] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Emomix: Emotion mixing via diffusion models for emotional speech synthesis," *ArXiv*, vol. abs/2306.00648, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258999280

[10] R. Chevi and A. F. Aji, "Daisy-tts: Simulating wider spectrum of emotions via prosody embedding decomposition," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267783078

[11] T. P. Kalyan, P. Rao, P. Jyothi, and P. Bhattacharyya, "Narrator or character: Voice modulation in an expressive multi-speaker tts," *INTERSPEECH 2023*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259855306

[12] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *ArXiv*, vol. abs/2106.06103, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235417304

[13] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *ArXiv*, vol. abs/2312.15185, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266551115

[14] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 10 745–10 759, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247451056

[15] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *ArXiv*, vol. abs/2212.04356, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252923993

[17] E. Troiano, S. Padó, and R. Klinger, "Emotion ratings: How intensity, annotation confidence and agreements are entangled," in *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232092844