Motivational Speech Synthesis

Abstract. *Motivational speech* has emerged as a popular audiovisual phenomenon within Western subcultures, conveying strategies and principles for individual success through expressive, high-energy delivery. The presented paper artistically explores methods for synthesizing the distinctive prosodic patterns inherent to motivational speech, while critically examining its sociocultural foundations. Drawing on recent advances in emotion-controllable text-to-speech (TTS) systems and speech emotion recognition (SER), we employ deep learning models and frameworks to replicate and analyze motivational speech. Within our proposed architecture, we introduce a one-dimensional *motivational factor*, enabling the control of motivational prosody according to intensity. Situated within broader discourses on self-optimization and meritocracy, *Motivational Speech Synthesis*[3] contributes to the field of emotional speech synthesis, while also prompting reflection on the societal values mediated in such narratives.

Keywords: text-to-speech (TTS) · speech emotion recognition (SER) · emotional speech synthesis · artistic research · motivational speech

1 Introduction

Within the increasing popularity of fitness and entrepreneurship in Western subcultures, video clips of so-called *motivational speech* received millions of views across social media. Usually, those audiovisual artifacts show excerpts from presentations or interviews of people, in most cases male business leaders, authors, and other influential figures—who narrate about optimal instructions, principles, and strategies for success. Paired with epic and emotional background music, these videos aim to act as a vehicle for self-motivation and goal pursuit. With a primary target group of men, success is often tied to wealthiness, professional growth, or appeal to women while the same is obstructed by characteristics such as weakness, fragility, or discontinuity. Through motivational speech, a listener's ultimate goal is to obtain and shape a mindset which ensures them to be on the right path for achievements. Motivational speech emerges as a phenomenon

[3] https://anonymous.4open.science/r/motivational-speech-synthesis-DB5A/README.md

in a society of self-optimization, inherent in the ethos of constant productivity, self-isolation, competition, and meritocracy.

Focusing on the human voice as the primary medium within this audiovisual subculture, its characteristic prosodic patterns play a decisive role in the appearance and perception of motivational speech. With *Motivational Speech Synthesis*, we therefore aim to

1. replicate those prosodic features by introducing a novel approach to control them through a linear scale
2. while creating a space for artistic reflection to extract the underlying attitudes of this subculture as a whole.

Correlating with the generalization process of one universal way to success, as well as the presence of an anticipated forward movement into a listener's future in motivational speech itself, we use machine learning techniques to average web-scraped motivational speech into a text-to-motivational-speech model adjustable with a one-dimensional *motivational factor*.

Our *motivational factor* enables fine-grained intensity control over motivational speech prosody during inference. By reducing dimensionality, we derive a mapping from a three-dimensional emotional representation of speech into a one-dimensional scale, ranging from 0 (low *motivational factor*) to 1 (high *motivational factor*).

Representing the promise of social mobility embodied by motivational speech subculture, our *motivational factor* aligns with attitudes like "The harder you work, the more you can get". Despite this emphasis on individual determination, OECD (Organisation for Economic Co-operation and Development) data suggests that income, education, and occupational status are still strongly shaped by one's family background [10]. *Motivational Speech Synthesis* addresses aspects of our work ethic and how we approach our goals and challenges in life, while raising questions on how we define "success" at all.

Motivated by this artistic goal and inspired by the realm of emotional speech synthesis with its ongoing efforts to reproduce speech with increasing emotional nuance and expression, we conceptualized and developed an approach for emotional controlled text-to-speech (TTS) generation.

Even though *Motivational Speech Synthesis* strives for artistic reflection, we want to emphasize, that this project does not aim to judge any person actually benefiting from motivational speech or similar phenomena. We don't expose or look at people consuming motivational speech, but rather focus on revealing underlying circumstances and attitudes of those narratives, which arrive as symptoms of a society driven by self-optimization, growth, and success.

2 Related Work

Recent advancements in emotion-aware text-to-speech and speech emotion recognition have significantly enhanced the field of emotional speech synthesis. Although many state-of-the-art models in open-source research—such as XTTS-

v2[4], MetaVoice[5], Parler-TTS [6], or StyleTTS 2 [7]—are capable of producing high-quality speech, few offer the ability to generate speech with specific emotional inflections. Although voice cloning has already reached a high level of sophistication, the integration of prosodic variation into TTS systems remains a critical step towards synthesized, human-like sounding speech. By incorporating emotional nuances, these systems can improve mimicking the subtleties of human expression, further minimizing the gap between artificial and natural speech.

The model proposed by Cho et al. [3] allows emotion intensity control along with style transfer, while EmoKnob [2] provides fine-grained emotion modulation using few-shot samples of arbitrary emotions. After comparing the previously mentioned architectures and TTS frameworks, EmoKnob was the most suitable solution for our purposes. By extending the voice cloning-based TTS model *MetaVoice*, the authors constructed a speaker representation space that enables systematic modelling and comparison of speaker characteristics. Here, an emotional embedding is created by extracting an emotion direction vector, obtained by computing the normalized difference between the emotional and neutral embeddings of the same speaker. Subsequently, this embedding is added to the speaker representation space.

In the domain of SER, emotions are primarily represented in two ways: as discrete categories (e.g., happy, sad, angry) [4] or as positions in a continuous emotion space usually defined by three dimensions: valence, arousal, and dominance. The scales within this 3D emotion model range from negative to positive emotions (valence), calm to stimulated emotions (arousal), and submissive to dominant emotions (dominance) [15]. As our proposed *motivational factor* does not fit into any of these discrete categories, but rather spans across this 3D space, our research focused on architectures that embed emotions in this continuous space.

One accessible model that explores the potential of transformer-based architectures for improving SER by embedding analyzed speech into a 3D emotional space is a fine-tuned version of wav2vec 2.0 [16]. Another approach we examined is emotion2vec [8], which provides a speech emotion representation model in a higher dimension in addition to a SER foundation model classifying emotions into discrete categories. Due to limited availability of labeled data for emotion recognition [5], both models use self-supervised learning frameworks. Here, a common approach involves using pretrained self-supervised models, such as wav2vec 2.0 [1], which are trained on large-scale speech datasets, and fine-tuning them for emotion recognition tasks [11]. This methodology allows overcoming data scarcity by utilizing the rich representations learned from vast amounts of unlabeled speech data, thereby improving the performance of SER systems.

[4] https://github.com/coqui-ai/TTS
[5] https://github.com/metavoiceio/metavoice-src

3 Method

3.1 Preprocessing

Motivational speeches on social media platforms like YouTube exhibit a consistent structure, accompanied by dramatic instrumental music. To capture the speech content of these videos, a multi-stage preprocessing pipeline (see Figure 1) is employed. After collecting audio data from multiple platforms dedicated to motivational content, the speech components are isolated using the music source separation algorithm Demucs [13].

Fig. 1. Overview of the data processing pipeline.

Once separated, the extracted speech undergoes further refinement, including speech enhancement with ai-coustics' proprietary model called *Lark*[6] and transcription via Whisper [12].

3.2 Model architecture

After careful evaluation of existing text-to-speech (TTS) models capable of emotional control over generated audio, we decided to base our architecture upon established approaches. Many contemporary models operate on higher-dimensional emotional representations, such as those produced by the aforementioned SER models [16,8] to generate emotionally expressive speech. We recognized that this characteristic allows for the implementation of our *motivational factor* without the necessity of developing an entirely new TTS architecture. Specifically, given that an appropriate dimensionality reduction method exists, higher-dimensional emotional representations can be mapped onto a one-dimensional *motivational factor*. This factor ranges continuously from 0, indicating a low motivational prosody intensity, to 1, indicating a high motivational prosody intensity. Subsequently, the derived *motivational factor* can serve indirectly as an input condition specified by the user during inference.

Once we defined the three-dimensional *VAD* space as our high-dimensional representation, we projected our motivational speech corpus onto it using the inference model proposed by Wagner et al. [16]. To represent our *motivational factor* as a single dimension, we therefore reduced these three dimensions into one by applying the UMAP (Uniform Manifold Approximation and Projection) [9] algorithm, resulting in the desired projection ranging from 0 to 1.

[6] https://developers.ai-coustics.com/documentation

Speaker Embedding Averaging and Selection In our approach, speaker embeddings—which encode stylistic characteristics of a selected speaker or reference audio within a high-dimensional feature space—are provided to the model during inference to guide the synthesis accordingly. By generating distinct speaker embeddings corresponding to discrete steps within a fitting range for the *motivational factor*, these embeddings can also be used indirectly to represent different *motivational factors*. During inference, an embedding nearest to the specified motivational input value is selected to guide the speech generation (see Figure 2). In our implementation, we adopted EmoKnob [2] as our TTS model and computed averaged speaker embeddings in increments of 0.05. For each increment, a representative speaker embedding was obtained by calculating the mean of the k-nearest neighbor (kNN=400) embeddings within the speaker embedding space.

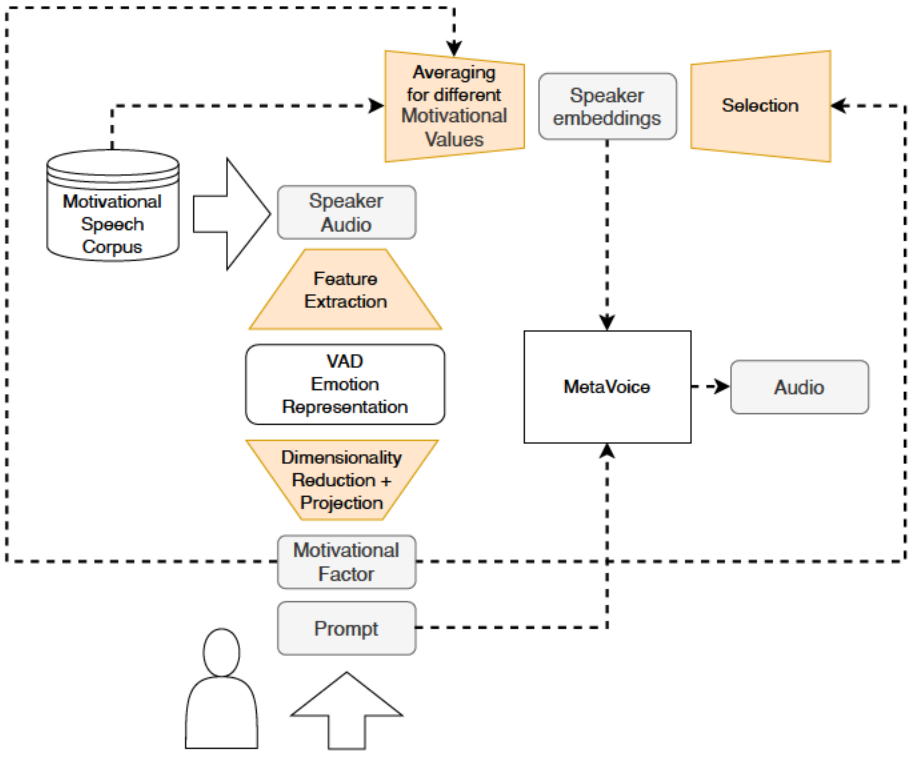

Fig. 2. Proposed model architecture with EmoKnob TTS model that uses *motivational factor* directly at inference and a reference audio averaged from the motivational speech corpus

4 Results

To synthesize motivational speech while being able to scale motivational prosody our EmoKnob based architecture involved generating averaged speaker embeddings, constituting a lightweight modification of the existing MetaVoice model without necessitating computationally intensive training or fine-tuning procedures.

During the development phase, we compiled an extensive motivational speech dataset comprising 414,024 data points with a total duration of approximately 371 hours.

To ensure a corpus of sufficient quality for speech synthesis, the preprocessing pipeline incorporated essential stages such as voice separation, speech enhancement, and transcription.

Figure 3 presents a visualization of a subset of $n = 2000$ audio data points, randomly selected along the *motivational factor* dimension, and projected into the VAD space. The resulting distribution reveals an arch-shaped trajectory, extending from regions of lower valence, arousal, and dominance toward higher arousal and dominance. This latent structure was effectively captured using the dimensionality reduction technique UMAP, supporting its suitability for representing the data along a single *motivational factor*.

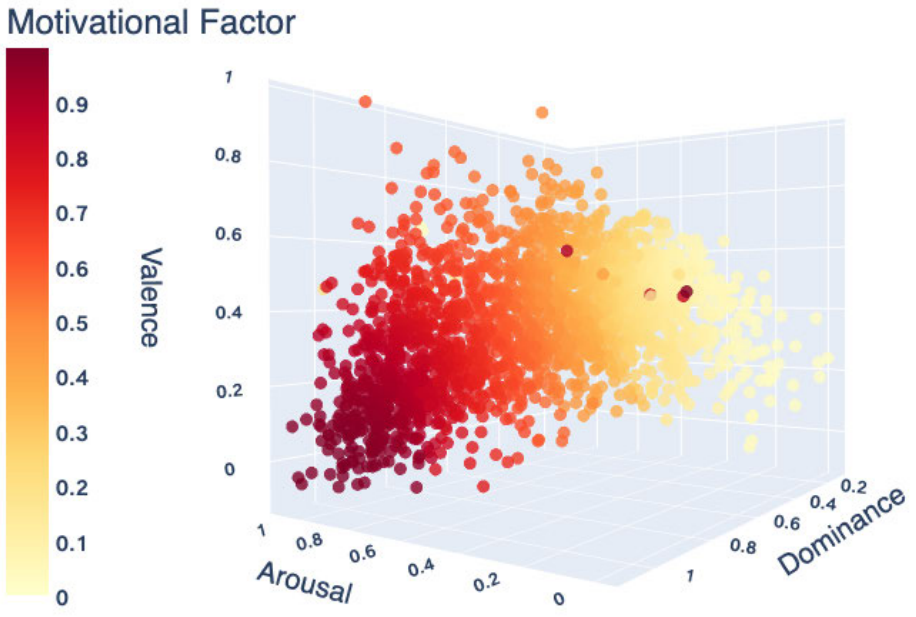

Fig. 3. Visualization of 2000 dataset audio points embedded into VAD space. *Motivational factor* representation via colormap.

While the main focus of our approach lay in achieving artistic results, we nevertheless included two quantitative evaluations to establish basic comparative benchmarks. First, our model achieved an average Word Error Rate (WER) of 0.21 utilizing Whisper [12] in its small version, which was being applied to motivational quotes of varied lengths. It should be noted that this WER value represents a conservative estimate, since it accounts for combined errors from both the synthesis and transcription model. Additionally, the naturalness of the synthesized audio was assessed using UTMOS [14] yielding an average Naturalness Mean Opinion Score (nMOS) of 3.22 out of 5, indicating fair perceived quality by human listeners.

5 Conclusion

Our research successfully introduced and evaluated a novel method for synthesizing motivational speech using averaged speaker embeddings within a modified EmoKnob architecture. By demonstrating the effective compression of dimensional emotional relationships into a singular motivational intensity scale, the developed method provides an intuitive control mechanism for adjusting motivational prosody in speech synthesis. While the synthesized outputs showed acceptable intelligibility and naturalness, several limitations were noted regarding transcription accuracy and audio quality.

Further analysis on how well the motivational prosody is captured in our motivational factor may be necessary to validate its perceptual relevance across diverse listener groups and application contexts. This includes exploring correlations between the motivational factor and human emotional cues, as well as testing its adaptability across different speaker identities and linguistic content.

We hope that our proposed architectures will inspire future research in the broader context of emotion-specific speech synthesis across various tasks and domains.

Acknowledgments. We gratefully acknowledge ai-coustics for providing their speech-enhancement model to improve our dataset, as well as the Berlin University of the Arts for supplying the computational resources.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: Advances in Neural Information Processing Systems. vol. 33, pp. 12449–12460. Curran Associates, Inc. (2020)
2. Chen, H., Chen, R., Hirschberg, J.: EmoKnob: Enhance Voice Cloning with Fine-Grained Emotion Control (Oct 2024). https://doi.org/10.48550/arXiv.2410.00316, http://arxiv.org/abs/2410.00316, arXiv:2410.00316 [cs]

3. Cho, D.H., Oh, H.S., Kim, S.B., Lee, S.W.: EmoSphere++: Emotion-Controllable Zero-Shot Text-to-Speech via Emotion-Adaptive Spherical Vector (Nov 2024). https://doi.org/10.48550/arXiv.2411.02625, http://arxiv.org/abs/2411.02625, arXiv:2411.02625 [cs]
4. Ekman, P.: An argument for basic emotions. Cognition & Emotion **6**(3-4), 169–200 (1992), publisher: Taylor & Francis
5. George, S.M., Muhamed Ilyas, P.: A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. Neurocomputing **568**, 127015 (Feb 2024). https://doi.org/10.1016/j.neucom.2023.127015, https://www.sciencedirect.com/science/article/pii/S0925231223011384
6. Lacombe, Y., Srivastav, V., Gandhi, S.: Parler-TTS (2024), https://github.com/huggingface/parler-tts, publication Title: GitHub repository
7. Li, Y.A., Han, C., Raghavan, V.S., Mischler, G., Mesgarani, N.: StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models (Nov 2023). https://doi.org/10.48550/arXiv.2306.07691, http://arxiv.org/abs/2306.07691, arXiv:2306.07691 [eess]
8. Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., Chen, X.: emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 15747–15760. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting (2024). https://doi.org/10.18653/v1/2024.findings-acl.931, https://aclanthology.org/2024.findings-acl.931
9. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction (2020), https://arxiv.org/abs/1802.03426
10. OECD: A Broken Social Elevator? How to Promote Social Mobility. OECD (Jun 2018). https://doi.org/10.1787/9789264301085-en, https://www.oecd.org/en/publications/broken-elevator-how-to-promote-social-mobility_9789264301085-en.html
11. Pepino, L., Riera, P., Ferrer, L.: Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings (Apr 2021). https://doi.org/10.48550/arXiv.2104.03502, http://arxiv.org/abs/2104.03502, arXiv:2104.03502 [cs]
12. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision (2022). https://doi.org/10.48550/ARXIV.2212.04356, https://arxiv.org/abs/2212.04356
13. Rouard, S., Massa, F., Défossez, A.: Hybrid Transformers for Music Source Separation. In: ICASSP 23 (2023)
14. Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., Saruwatari, H.: UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022 (2022), https://arxiv.org/abs/2204.02152, _eprint: 2204.02152
15. Verma, G.K., Tiwary, U.S.: Affect representation and recognition in 3D continuous valence–arousal–dominance space. Multimedia Tools and Applications **76**(2), 2159–2183 (Jan 2017). https://doi.org/10.1007/s11042-015-3119-y, https://doi.org/10.1007/s11042-015-3119-y
16. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., Schuller, B.W.: Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10745–10759 (Sep 2023). https://doi.org/10.1109/TPAMI.2023.3263585, https://ieeexplore.ieee.org/document/10089511/