Towards Motivational Speech Synthesis

Luis Küffner
Art and Media
Berlin University of Arts
Berlin, Germany
l.kueffner@udk-berlin.de

Pierre-Louis Suckrow
Design and Computation
Berlin University of Arts
Berlin, Germany
p.suckrow@udk-berlin.de

Patrick Stecher
Computer Science
Technical University Berlin
Berlin, Germany
stecher@campus.tu-berlin.de

Nikolaj Wolff
Audiocommunication and -technology
Technical University Berlin
Berlin, Germany
n.wolff@campus.tu-berlin.de

***Abstract*—*Motivational speech* has emerged as a popular audiovisual phenomenon within Western subcultures, conveying optimal strategies and principles for success through expressive, high-energy delivery. The present 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 this genre of speech. Within our proposed architecture, we introduce a one-dimensional *motivational factor* derived from high-dimensional emotional speech representations, enabling the control of motivational prosody according to intensity. Situated within broader discourses on self-optimization and meritocracy, *Motivational Speech Synthesis* contributes to the field of emotional speech synthesis, while also prompting reflection on the societal values embedded in such mediated narratives[1].**

***Index Terms*—text-to-speech (TTS), speech emotion recognition (SER), emotional speech synthesis, motivational speech, artistic research**

I. 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 achievement. Motivational speech emerges as a phenomenon in a society of self-optimization, embedded 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 specific prosodic features
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*.

This concept of a *motivational factor* achieves fine-grained intensity control over motivational speech prosody during inference. Utilizing dimensionality reduction methods, 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*). Consequently, the validity and applicability of capturing motivational speech prosody through this dimensional compression prompts our first research question: **RQ1:** Can higher-dimensional emotional relationships in speech be effectively compressed into a singular one-dimensional scale representing motivational intensity?

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 data suggests that income, education, and occupational status are still strongly shaped by one's family background [11]. *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 explored and conceptualized different approaches for emotional controlled text-to-speech (TTS) generation. Spanning across different machine learning frameworks and speech

[1]https://github.com/MotivationalSpeechSynthesis/motivational-speech-synthesis

emotion recognition (SER) systems, we present multiple implementation possibilities as well as one realized motivational TTS architecture.

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 deconstructing underlying circumstances and attitudes of those narratives, which arrive as symptoms of a society driven by growth and success.

II. 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—such as XTTS-v2[2], MetaVoice[3], Parler-TTS [7], or StyleTTS 2 [9]—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 human speech.

The model proposed by Cho et al. [4] 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 building on the voice cloning-based TTS model MetaVoice, the authors established a speaker representation space. Here, an emotional embedding is created by calculating the difference between an emotional sample and a corresponding neutral sample, both spoken by 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) [5] 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) [17]. 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 by Wagner et al. [18]. Another approach we examined is emotion2vec [10], 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 [6], 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 [12]. 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.

III. Method

A. Preprocessing

Motivational speeches on social media platforms like YouTube exhibit a consistent structure, typically comprising curated excerpts from coaches, public figures (e.g., actors or professional athletes), 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 YouTube channels dedicated to motivational content, the speech components are isolated using the music source separation algorithm Demucs [15].

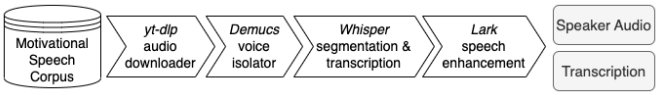

Fig. 1. Overview of the data processing pipeline.

Once separated, the extracted speech undergoes further refinement, including speech enhancement with ai|coustics' proprietary model called *Lark*[4] and transcription via Whisper [14].

B. 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 [18, 10] 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 state, to 1, indicating a high motivational state.

[2]https://github.com/coqui-ai/TTS

[3]https://github.com/metavoiceio/metavoice-src

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

Subsequently, the derived *motivational factor* can serve directly—or indirectly, by mapping it to a higher dimension—as a conditioning parameter during model training or 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. [18]. To represent our *motivational factor* as a single dimension, we therefore reduced these three dimensions into one by applying the UMAP algorithm, resulting in the desired projection ranging from 0 to 1.

We propose three distinct methodological approaches for integrating the *motivational factor* into existing TTS architectures, including a concrete implementation based on the EmoKnob framework.

1) Dimensional emotion conditioning: Multiple systems proposed by Li and Chen [8], Qi et al. [13], and Cho et al. [3, 4] aim to achieve controllable emotions in TTS generation by using pretrained SER frameworks, within their architectures. In our approach, this serves as a foundation for indirectly controlling the desired *motivational factor* by learning or defining an inverse mapping from its one-dimensional value to a higher-dimensional emotional representation. The resulting values can then serve as conditioning inputs during inference, enabling speech synthesis that reflects the intended style (see Figure 2).

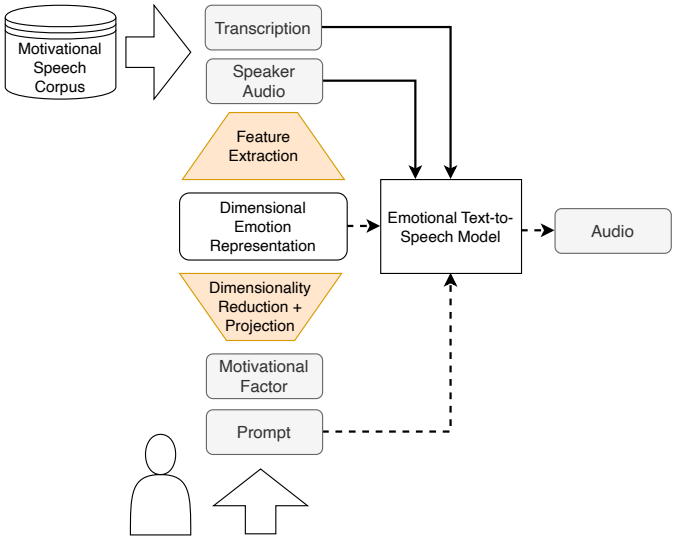

Fig. 2. Proposed general model architecture with dimensional emotion conditioning. Dashed lines represent inference, solid lines training.

2) Reference Audio Selection: Other models, such as XTTS v2, enable guidance during inference through the use of reference audio, often used for voice cloning. By selecting reference audio corresponding to the given value, this allows us to model the *motivational factor* indirectly (see Figure 3). In combination with fine-tuning the model on our motivational speech corpus, this approach enables single-speaker synthesis with averaged motivational prosody. Furthermore, a selection algorithm can be designed to introduce controlled variability by randomly choosing different reference audio samples corresponding to a given *motivational factor* from the dataset. Simultaneously, consistency can be achieved by reusing selected reference audio samples across multiple generation tasks.

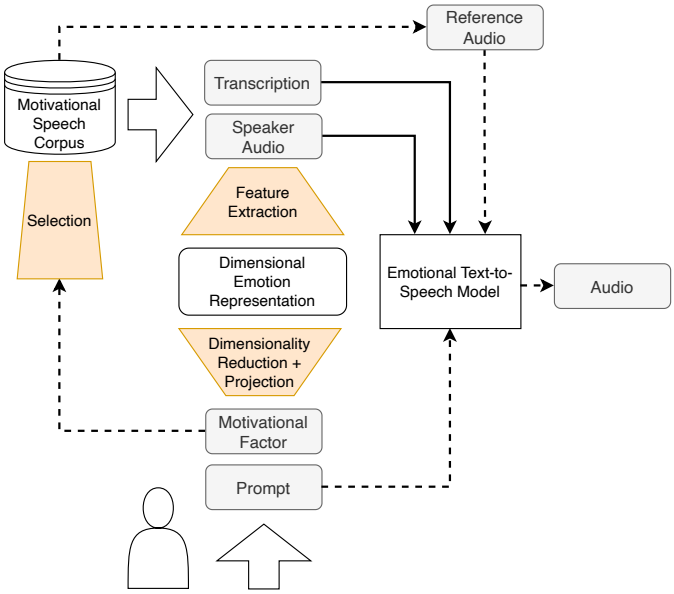

Fig. 3. Proposed model architecture with TTS model that takes a reference audio. Dashed lines represent inference, solid lines training.

3) Speaker Embedding Averaging and Selection: In this 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 production (see Figure 4). 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.

IV. Results

Among those three proposed methods for synthesizing motivational speech, we implemented the approach based on EmoKnob, as detailed in section III-B3. Our chosen architecture involved generating averaged speaker embeddings, constituting a lightweight modification of the existing MetaVoice

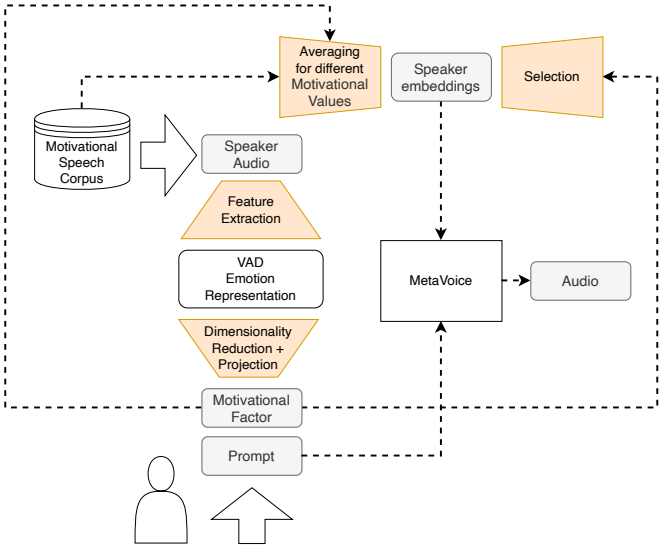

Fig. 4. Proposed model architecture with EmoKnob TTS model that uses *motivational factor* directly at inference and a reference audio averaged from the motivational speech corpus

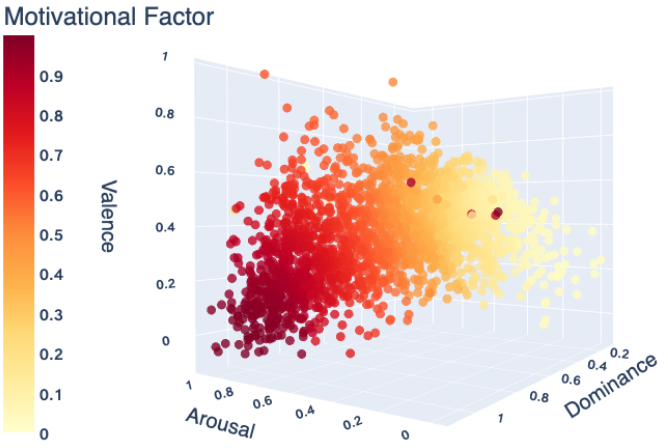

Fig. 5. Visualization of 2000 dataset audio points embedded into VAD space. *Motivational factor* representation via colormap.

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 5 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*.

The speech synthesized by EmoKnob was quantitatively evaluated using two metrics. First, the model achieved an average Word Error Rate (WER) of 0.21 utilizing Whisper [14] 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 [16] yielding an average Naturalness Mean Opinion Score (nMOS) of 3.22 out of 5, indicating fair perceived quality by human listeners. The second proposed approach, involving high-dimensional emotional conditioning (section III-B1), could not be practically evaluated due to limitations in available models and architectures.

V. 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. Additionally, practical barriers encountered in implementing alternative high-dimensional emotional conditioning approaches highlight the necessity for improving the accessibility and maintainability of computational resources in emotional speech synthesis.

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 contribute to future research not only in the modeling of motivational speech, but also in the broader context of emotion-specific speech synthesis across various tasks and domains.

References

[1] Alexei Baevski et al. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 12449–12460.

[2] Haozhe Chen, Run Chen, and Julia Hirschberg. *EmoKnob: Enhance Voice Cloning with Fine-Grained Emotion Control*. en. arXiv:2410.00316 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2410.00316.

[3] Deok-Hyeon Cho et al. "EmoSphere-TTS: Emotional Style and Intensity Modeling via Spherical Emotion Vector for Controllable Emotional Text-to-Speech". In: *Interspeech 2024*. arXiv:2406.07803 [cs]. Sept. 2024, pp. 1810–1814. DOI: 10.21437/Interspeech.2024-398.

[4] Deok-Hyeon Cho et al. *EmoSphere++: Emotion-Controllable Zero-Shot Text-to-Speech via Emotion-Adaptive Spherical Vector*. arXiv:2411.02625 [cs]. Nov. 2024. DOI: 10.48550/arXiv.2411.02625.

[5] Paul Ekman. "An argument for basic emotions". In: *Cognition & Emotion* 6.3-4 (1992). Publisher: Taylor & Francis, pp. 169–200.

[6] Swapna Mol George and P. Muhamed Ilyas. "A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise". In: *Neurocomputing* 568 (Feb. 2024), p. 127015. DOI: 10.1016/j.neucom.2023.127015.

[7] Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. *Parler-TTS*. Publication Title: GitHub repository. 2024.

[8] Guopping Li and Yanxiang Chen. "Intensity Controllable Emotional Speech Synthesis Based on Valence-Arousal-Dominance". en. In: *Advances in Brain Inspired Cognitive Systems*. Ed. by Amir Hussain et al. Singapore: Springer Nature, 2025, pp. 30–40. ISBN: 978-981-9628-82-7. DOI: 10.1007/978-981-96-2882-7_4.

[9] Yinghao Aaron Li et al. *StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models*. arXiv:2306.07691 [eess]. Nov. 2023. DOI: 10.48550/arXiv.2306.07691.

[10] Ziyang Ma et al. "emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation". en. In: *Findings of the Association for Computational Linguistics ACL 2024*. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, 2024, pp. 15747–15760. DOI: 10.18653/v1/2024.findings-acl.931.

[11] OECD. *A Broken Social Elevator? How to Promote Social Mobility*. en. OECD, June 2018. ISBN: 978-92-64-30107-8 978-92-64-30108-5. DOI: 10.1787/9789264301085-en.

[12] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. *Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings*. arXiv:2104.03502 [cs]. Apr. 2021. DOI: 10.48550/arXiv.2104.03502.

[13] Tianhua Qi et al. *Towards Realistic Emotional Voice Conversion using Controllable Emotional Intensity*. arXiv:2407.14800 [eess]. July 2024. DOI: 10.48550/arXiv.2407.14800.

[14] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. DOI: 10.48550/ARXIV.2212.04356.

[15] Simon Rouard, Francisco Massa, and Alexandre Défossez. "Hybrid Transformers for Music Source Separation". In: *ICASSP 23*. 2023.

[16] Takaaki Saeki et al. *UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022*. _eprint: 2204.02152. 2022.

[17] Gyanendra K. Verma and Uma Shanker Tiwary. "Affect representation and recognition in 3D continuous valence–arousal–dominance space". en. In: *Multimedia Tools and Applications* 76.2 (Jan. 2017), pp. 2159–2183. ISSN: 1573-7721. DOI: 10.1007/s11042-015-3119-y.

[18] Johannes Wagner et al. "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (Sept. 2023), pp. 10745–10759. DOI: 10.1109/TPAMI.2023.3263585.