Towards Motivational Speech Synthesis

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]

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 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 artistically reflecting and extracting underlying attitudes of this subculture at large.

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

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

(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 within the 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 (OECD, 2018). *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 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.

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 - such as XTTS-v2[2], MetaVoice[3], Parler-TTS (Lacombe et al., 2024), or StyleTTS 2 (Li et al., 2023) — 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, natural and 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. (2024b) allows emotion intensity control along with style transfer, while EmoKnob (Chen et al., 2024) 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, they 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) (Ekman, 1992) 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) (Verma and Tiwary, 2017). 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. (2023). Another approach we examined is emotion2vec (Ma et al., 2024), 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

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

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

data for emotion recognition (George and Muhamed Ilyas, 2024), both models use self-supervised learning frameworks. Here, a common approach involves using pretrained self-supervised models, such as wav2vec 2.0 (Baevski et al., 2020), which are trained on large-scale speech datasets, and fine-tuning them for emotion recognition tasks (Pepino et al., 2021). 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.

3 Method

3.1 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. First, audio data is collected from multiple YouTube channels dedicated to motivational content. The speech component is then isolated using the music source separation algorithm Demucs (Rouard et al., 2023).

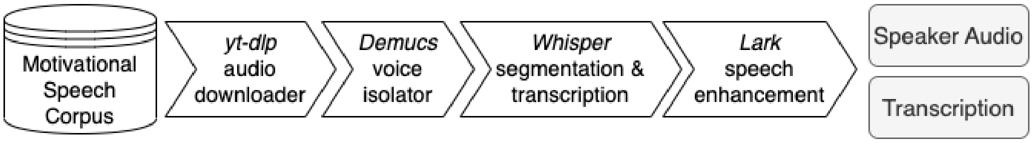
Motivational Speech Corpus
yt-dlp audio downloader
Demucs voice isolator
Whisper segmentation & transcription
Lark speech enhancement
Speaker Audio
Transcription

Figure 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 is transcribed via Whisper (Radford et al., 2022).

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 (Wagner et al., 2023; Ma et al., 2024) 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. 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.

We selected the three-dimensional *VAD* space as our higher-dimensional representation and employed an inference model by Wagner et al. (2023) to convert our motivational speech corpus into this space. To represent our *motivational factor* as a single dimension, we reduced these three dimensions into one by applying the UMAP algorithm, resulting in the wanted projection ranging from 0 to 1.

We propose three distinct methodological approaches for integrating the *motivational factor* around available TTS-architecture, with one concrete implementation example around the EmoKnob framework.

3.2.1 Dimensional emotion conditioning

Multiple approaches proposed by Li and Chen (2025); Qi et al. (2024); Cho et al. (2024a,b) aim to achieve controllable emotions in TTS generation by using pretrained SER frameworks, within

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

their architectures. This provides a starting point to indirectly control the desired *motivational factor*. Specifically, by learning or defining an inverse mapping from the one-dimensional *motivational factor* to the higher-dimensional emotional representation, the resulting value can potentially serve at inference as conditioning input, enabling the generation of speech with a style corresponding to the specified *motivational factor* (see figure 2).

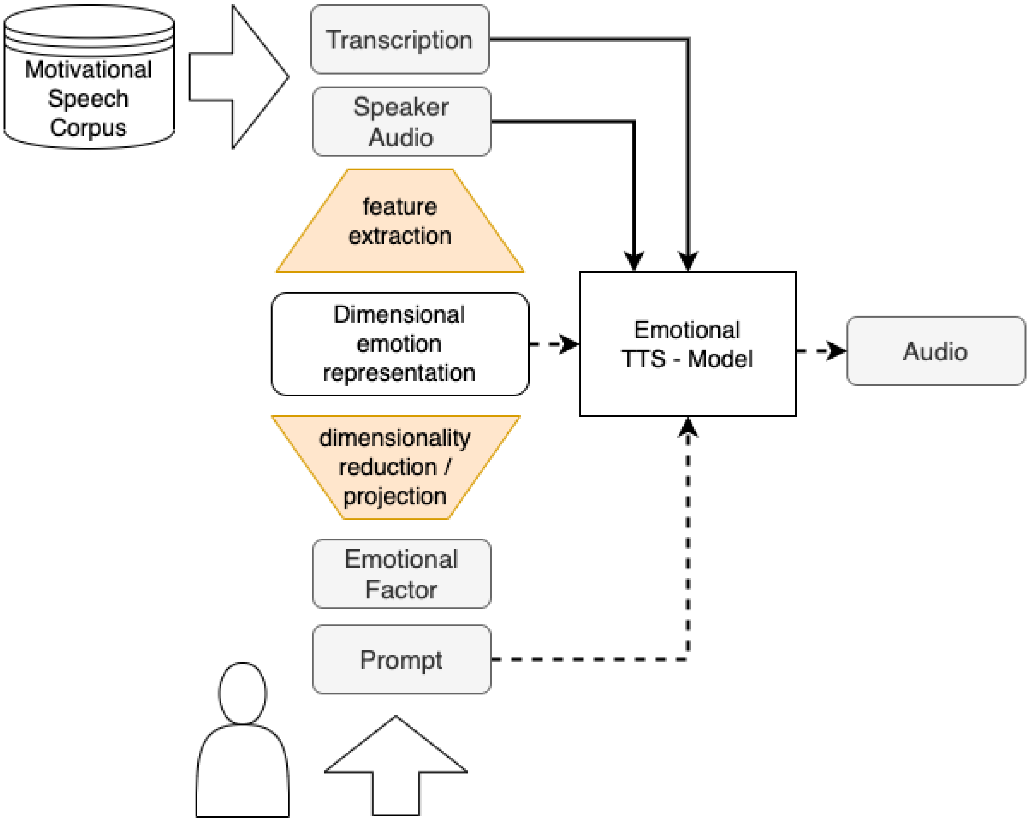
Motivational Speech Corpus
Transcription
Speaker Audio
feature extraction
Dimensional emotion representation
dimensionality reduction / projection
Emotional Factor
Prompt
Emotional TTS - Model
Audio

Figure 2: Proposed general model architecture with dimensional emotion conditioning. Dashed line representing inference and the solid line training.

3.2.2 Reference Audio Selection

Other models, such as XTTS v2, enable guidance during inference through the use of reference audio, mainly used for voice cloning. This allows the indirect modeling of our *motivational factor* by selecting reference audio corresponding to the given value (see figure 3). In combination with fine-tuning the model on our motivational speech corpus, this approach supports the synthesis of averaged motivational characteristics in a single speaker. Furthermore, the 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.

3.2.3 Speaker Embedding Averaging and Selection

Speaker embeddings represent the style characteristics of the selected speaker or reference audio in a high-dimensional feature space, which is passed to the model during inference to be applied during generation. These speaker embeddings can also be indirectly used to represent *motivational factor*s by generating distinct embeddings corresponding to discrete steps within a *motivational factor* range.

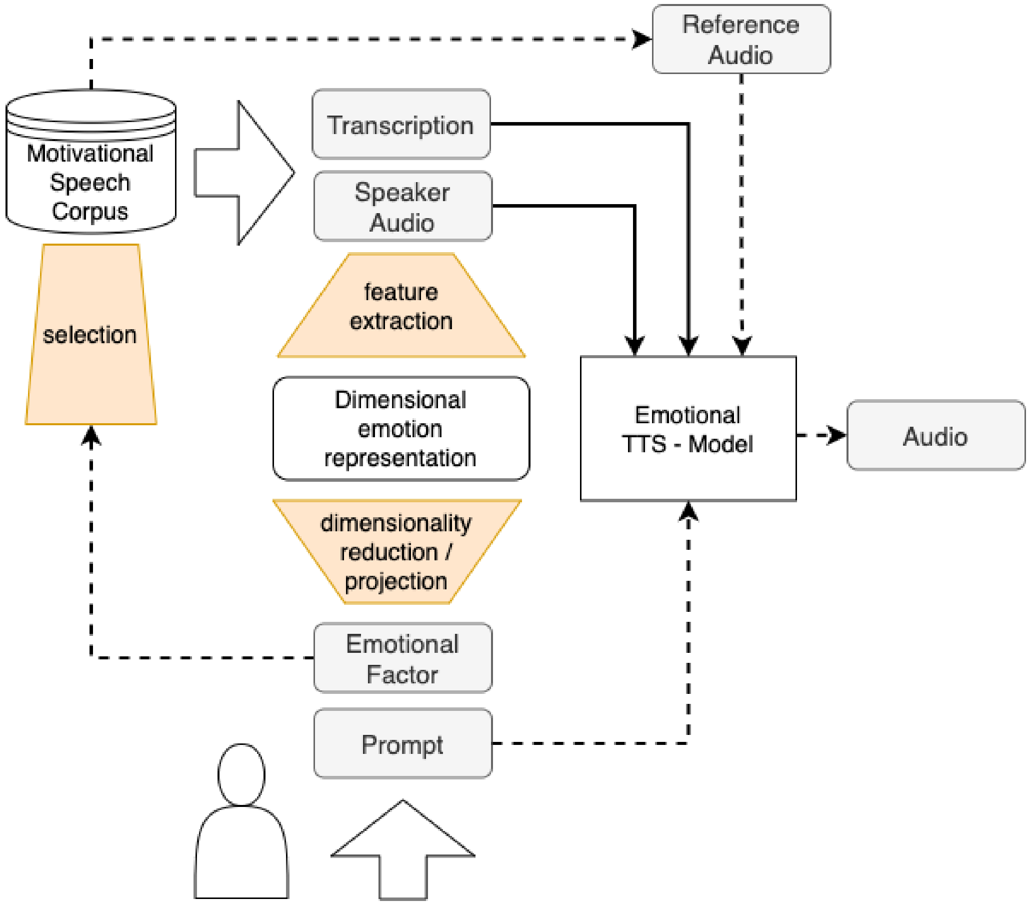

Figure 3: Proposed model architecture with TTS model that takes a reference audio. Dashed line representing inference and the solid line training.

During inference, the embedding nearest to the specified motivational input value is selected and subsequently applied to guide the audio synthesis process (see figure 4)

For our use, we adopted EmoKnob (Chen et al., 2024) as our TTS model and prepared speaker embeddings in increments of 0.05. For each increment, we took the mean of the k-nearest neighbors (kNN) speaker embeddings, resulting in averaged speaker embeddings.

4 Results

Among the three approaches proposed for modeling motivational speech synthesis, we implemented the methodology based on EmoKnob, as detailed in section 3.2.3. Our chosen 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 dataset of motivational speech. This dataset comprised 414.024 data points and had a cumulative duration of approx. 371 hours. The preprocessing pipeline included several essential steps, such as voice separation, audio enhancement, and transcription, ensuring a high-quality corpus suitable for synthesis.

Figure 5 displays a visualization of a sample containing $n = 2000$ audio data points—selected randomly along the *motivational factor* dimension—within the VAD (valence-arousal-dominance) space, revealing an arch-shaped distribution. Specifically, the data points extended from lower valence, arousal, and dominance regions toward higher arousal and dominance areas. This underlying

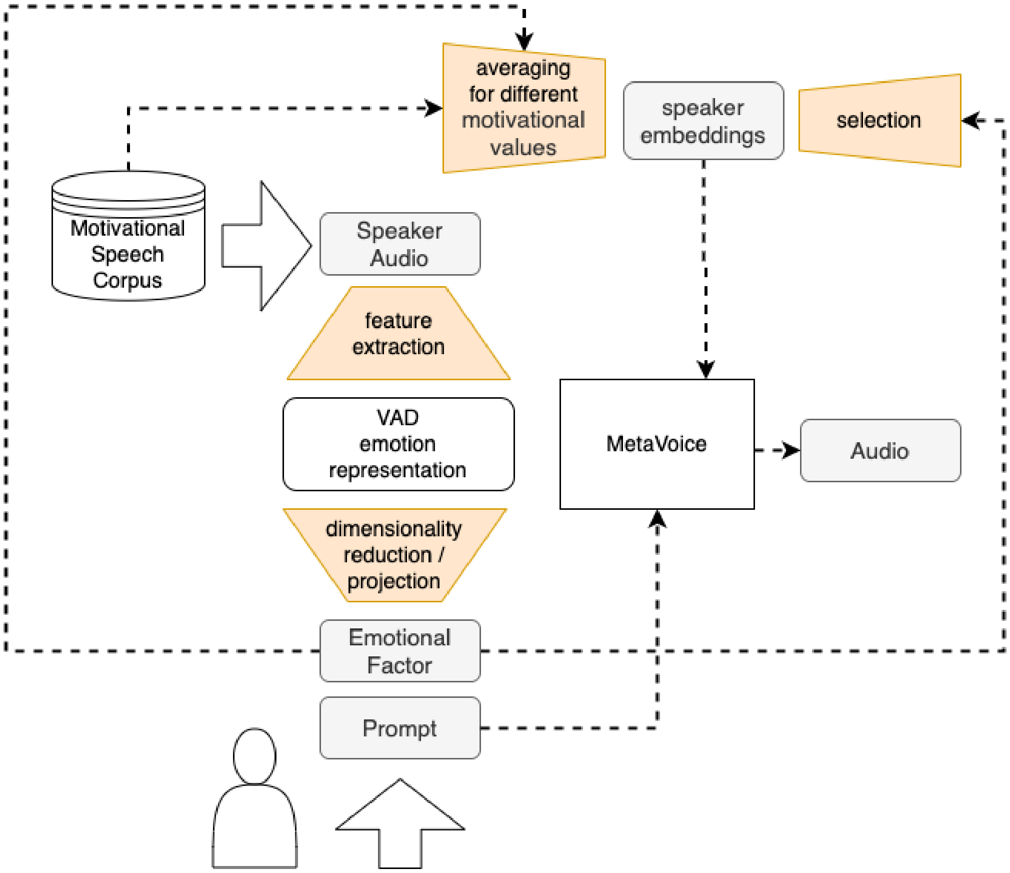
averaging for different motivational values
speaker embeddings
selection
Motivational Speech Corpus
Speaker Audio
feature extraction
VAD emotion representation
MetaVoice
Audio
dimensionality reduction / projection
Emotional Factor
Prompt

Figure 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

structure was effectively captured by our utilized dimensionality reduction method UMAP, enabling representation via 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 (Radford et al., 2022) 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 (Saeki et al., 2022) 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 3.2.1), could not be practically evaluated due to limitations in available models and architectures.

5 Conclusion

This study 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

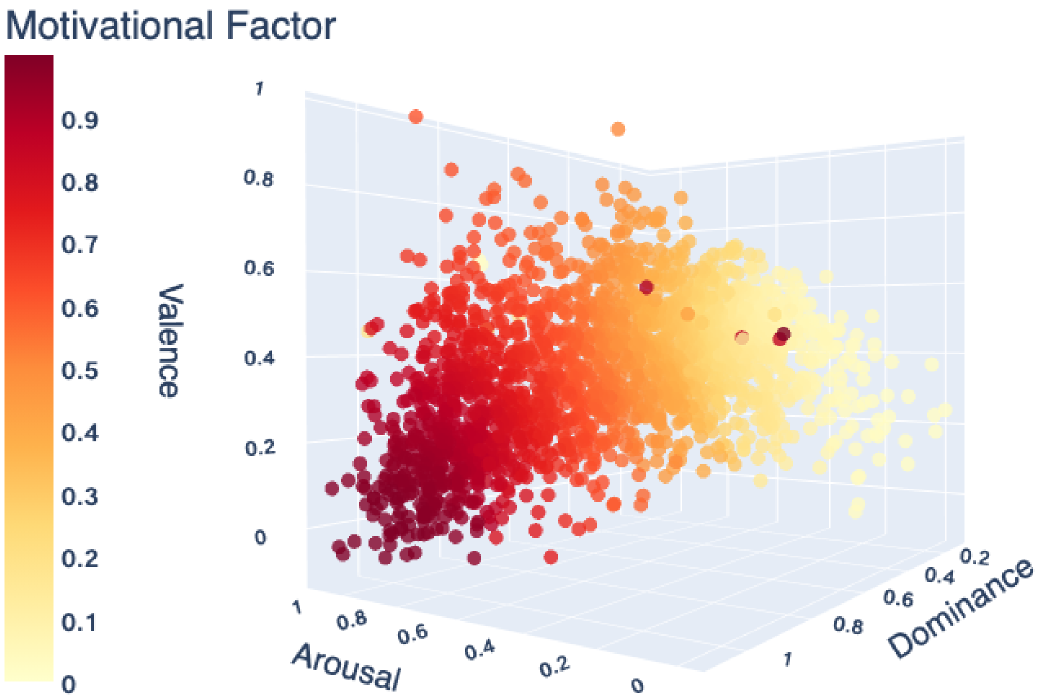

Figure 5: Visualization of 2000 dataset audio points embedded into VAD space. *Motivational factor* representation via colormap.

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

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Chen, H., Chen, R., and Hirschberg, J. (2024). EmoKnob: Enhance Voice Cloning with Fine-Grained Emotion Control. arXiv:2410.00316 [cs].

Cho, D.-H., Oh, H.-S., Kim, S.-B., Lee, S.-H., and Lee, S.-W. (2024a). EmoSphere-TTS: Emotional Style and Intensity Modeling via Spherical Emotion Vector for Controllable Emotional Text-to-Speech. In *Interspeech 2024*, pages 1810–1814. arXiv:2406.07803 [cs].

Cho, D.-H., Oh, H.-S., Kim, S.-B., and Lee, S.-W. (2024b). EmoSphere++: Emotion-Controllable Zero-Shot Text-to-Speech via Emotion-Adaptive Spherical Vector. arXiv:2411.02625 [cs].

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200. Publisher: Taylor & Francis.

George, S. M. and Muhamed Ilyas, P. (2024). A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568:127015.

Lacombe, Y., Srivastav, V., and Gandhi, S. (2024). Parler-TTS. Publication Title: GitHub repository.

Li, G. and Chen, Y. (2025). Intensity Controllable Emotional Speech Synthesis Based on Valence-Arousal-Dominance. In Hussain, A., Jiang, B., Ren, J., Mahmud, M., Yang, E., Zheng, A., Li, C., Wang, S., Gao, Z., and Zhao, Z., editors, *Advances in Brain Inspired Cognitive Systems*, pages 30–40, Singapore. Springer Nature.

Li, Y. A., Han, C., Raghavan, V. S., Mischler, G., and Mesgarani, N. (2023). StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. arXiv:2306.07691 [eess].

Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., and Chen, X. (2024). emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15747–15760, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

OECD (2018). *A Broken Social Elevator? How to Promote Social Mobility*. OECD.

Pepino, L., Riera, P., and Ferrer, L. (2021). Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. arXiv:2104.03502 [cs].

Qi, T., Wang, S., Lu, C., Zhao, Y., Zong, Y., and Zheng, W. (2024). Towards Realistic Emotional Voice Conversion using Controllable Emotional Intensity. arXiv:2407.14800 [eess].

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.

Rouard, S., Massa, F., and Défossez, A. (2023). Hybrid Transformers for Music Source Separation. In *ICASSP 23*.

Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., and Saruwatari, H. (2022). UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. _eprint: 2204.02152.

Verma, G. K. and Tiwary, U. S. (2017). Affect representation and recognition in 3D continuous valence–arousal–dominance space. *Multimedia Tools and Applications*, 76(2):2159–2183.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., and Schuller, B. W. (2023). 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.