Style-Controlled VALL-E for Few-Shot Emotional German TTS

Rami Kammoun, **Mohammed Salah Al-Radhi**

malradhi@tmit.bme.hu

October 21, 2025

Background

Neural Codec Models in TTS

Emotion is the missing layer between intelligibility (what is said) and authenticity (how it is felt).

- Most emotional TTS systems require thousands of samples per emotion → low-resource languages (ex., German) are limited to a few hours.
- Traditional TTS → predict acoustic features, vocode waveforms
- Neural codecs (VALL-E, EnCodec) → model speech as discrete tokens
  - Enable zero-shot voice cloning and style transfer

Emotional and low-resource languages still underexplored

Why Emotional TTS for German?

- No large-scale emotional German datasets (LibriTTS equivalents don’t exist).
- English models train on ≈60,000 hours. For German, we have <3 hours.
- Multilingual transfer often loses expressivity and emotion.

- Few-shot learning can adapt models with ≈ 3 hours of data.

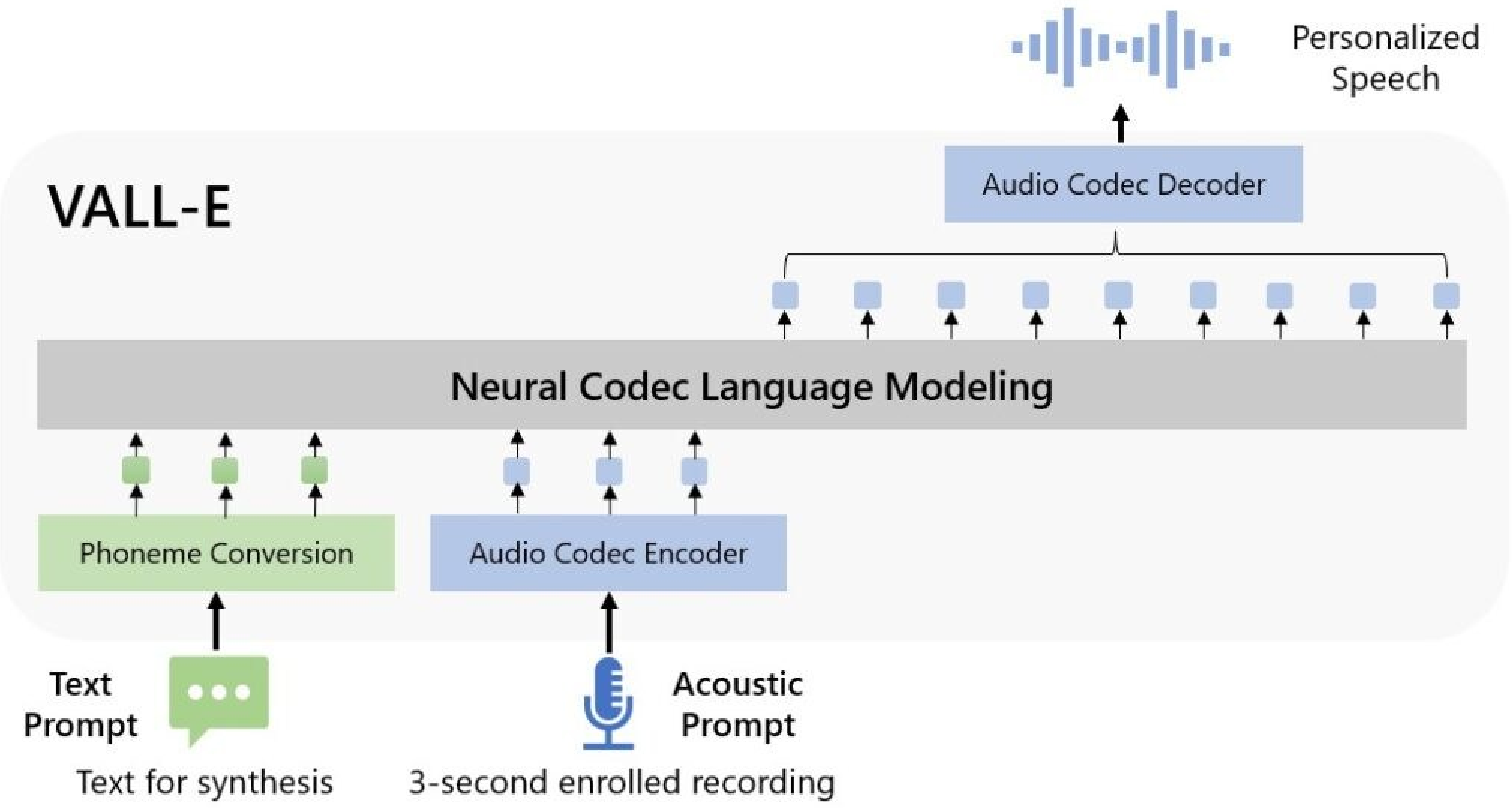
Goal: Controllable emotional TTS for low-resource German speech.

Related Work

Model	Emotion Control	Few-Shot Style	Low-Resource Language	Notes
Tacotron2	❌	Partial	EN	High-fidelity, but lacks explicit style/emotion axis
FastSpeech2	🔮	Limited	EN	Speed improvement comes at a cost of prosodic richness
VITS	🔮	Partial	EN/CH	Good naturalness, but still relies on large, labeled data
SC-VALL-E	✅	No	EN	Style tokens for English. Fails the low-resource/German test
Our Model	✅	✅	✅ (DE)	First VALL-E-based solution to address all three constraints for German

VALL-E as a Generative Neural Codec LM

- VALL-E redefines TTS as an autoregressive language modeling task, not waveform or Mel-spectrogram synthesis.
- Uses discrete EnCodec tokens (e.g., 8-12 codebooks) for efficient sequence representation.
- Learns speaker identity, prosody, and context by predicting speech tokens based on a 3-second acoustic prompt.
- The tokenized approach enables in-context learning and zero/few-shot style transfer, making it inherently more data-efficient than continuous-feature models.



https://www.microsoft.com/en-us/research/project/vall-e-x/vall-e/

Methodology

Dataset Overview

- **Corpus:** SLR110 (German Emotional)
  - Publicly available benchmark for replication
- **Size:** ≈175 minutes (2,400 utterances)
  - Confirms the few-shot requirement
- **Emotions:** 8 Categories
  - (Neutral, Angry, Amused, Disgusted, Drunk, Sleepy, Surprised, Whispering)
- Identical sentences across emotions
- **Augmentation Trick:** Different emotion renditions of the same text are paired in training batches
  - Prevents linguistic features from leaking into the emotion/style embeddings

Pre-Processing & Data Preparation Pipeline

Linguistic Normalization

- Unicode normalization & special-character mapping (ä→ae, ß→ss, ö→oe)
- Lowercasing and punctuation filtering

Phoneme-Level Conversion

- Grapheme-to-Phoneme (G2P) mapping using eSpeak-NG

Acoustic Preprocessing

- Silence trimming via energy thresholding (< –35 dB)
- Time-stretch (±5 %) and pitch-shift (±2 semitones) augmentation

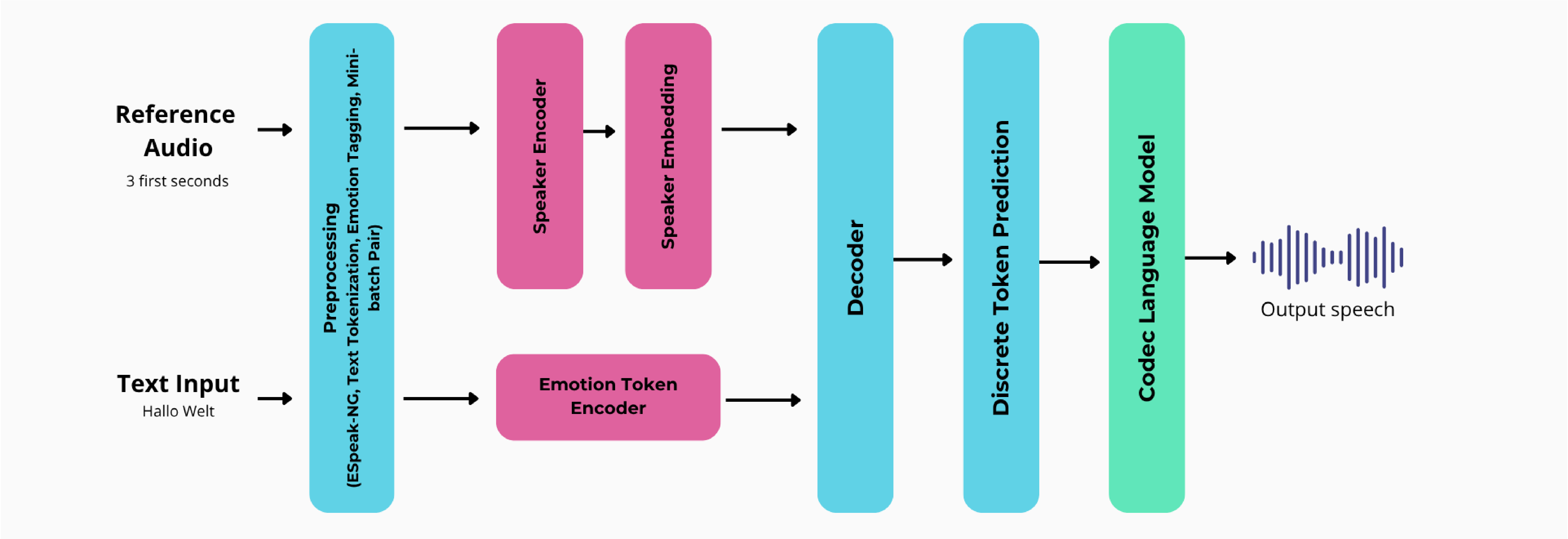
Dataset Structuring

- Split: Train 80 % , Validation 10 % , Test 10 %
- Metadata: transcript, emotion label, duration, speaker ID

Architecture Overview: Style-Controlled VALL-E

- Added branches:
  - Style Token Network → learns latent prosody.
  - Emotion Token Embedding → provides explicit emotion control.

- Decoder predicts discrete EnCodec tokens



Text → Phonemes → VALL-E Encoder → (Style + Emotion tokens) → Decoder → EnCodec

Style Token Network

- 16 learnable style tokens (256-D each), similar to a codebook of prosodic patterns.
- A reference encoder (e.g., CNN or Transformer) extracts global prosodic features from the 3s audio prompt.
- A style attention mechanism selects the most relevant of the 16 tokens to represent the reference's global style.
- The selected style embedding is injected into the autoregressive decoder layers (e.g., via Cross-Attention).

Emotion Token Conditioning

- 8 fixed, trained emotion embeddings (e.g., Neutral, Angry, Sleepy, etc.).
- The chosen emotion embedding is prepended to the phoneme token sequence, directly conditioning the autoregressive generation from the start.
- At inference, a single token flip (e.g., replacing 'Neutral' with 'Angry’) explicitly controls the output emotion while holding the style (from the 3s prompt) constant.

RESULTS

Experimental Setup and Insights

Parameter	Value/Setting	Technical Insight
Optimizer	Adam, LR = 3e−4	Lower Learning Rate was necessary to stabilize codebook learning.
Architecture	12-layer decoder, Dropout 0.1	Heavy regularization necessary to prevent overfitting on 175 minutes of data.
Critical Fix 1	Codebook Collapse: Occurred for LR ≥1e−3.	VALL-E's codebook space is highly sensitive; requires careful warmup and tuning.
Critical Fix 2	Unstable Prosody: Occurred for prompt length <2s.	Standardized reference prompt length to 3 seconds for reliable style embedding extraction.
Training Time	≈20 hours on RTX 4090	Fast convergence highlights the data efficiency of the neural codec approach.

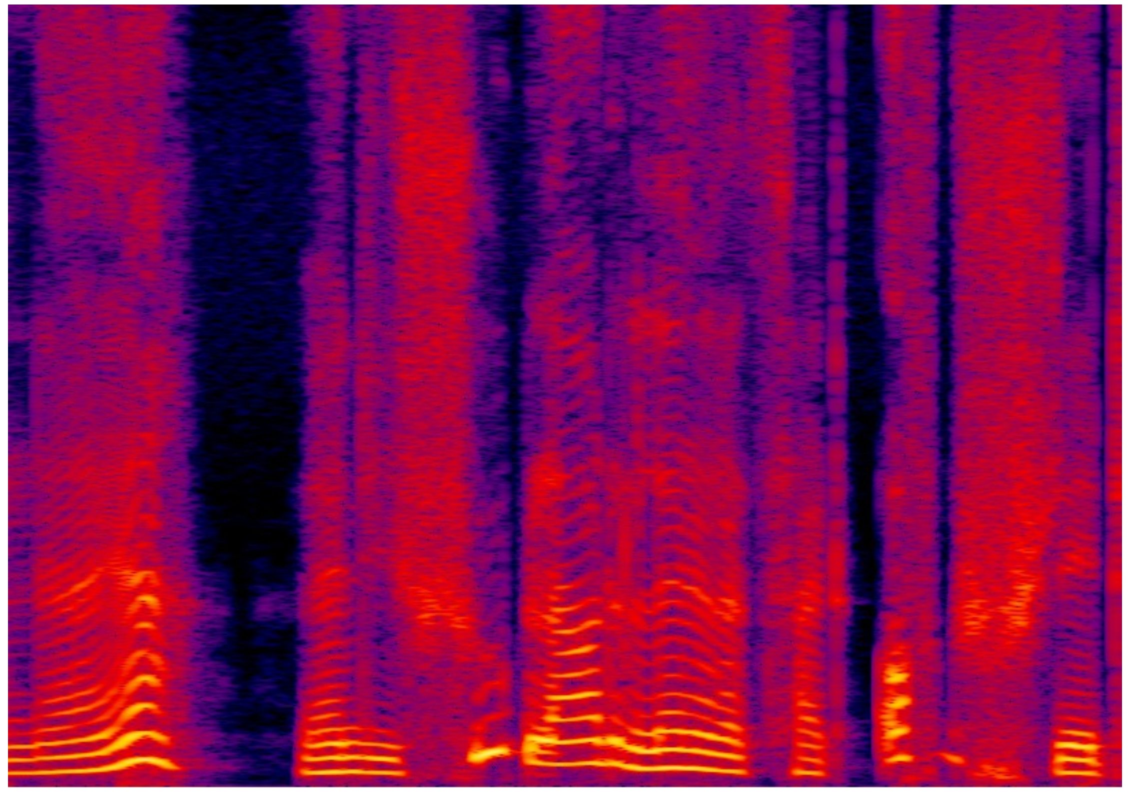
Evaluation Metrics

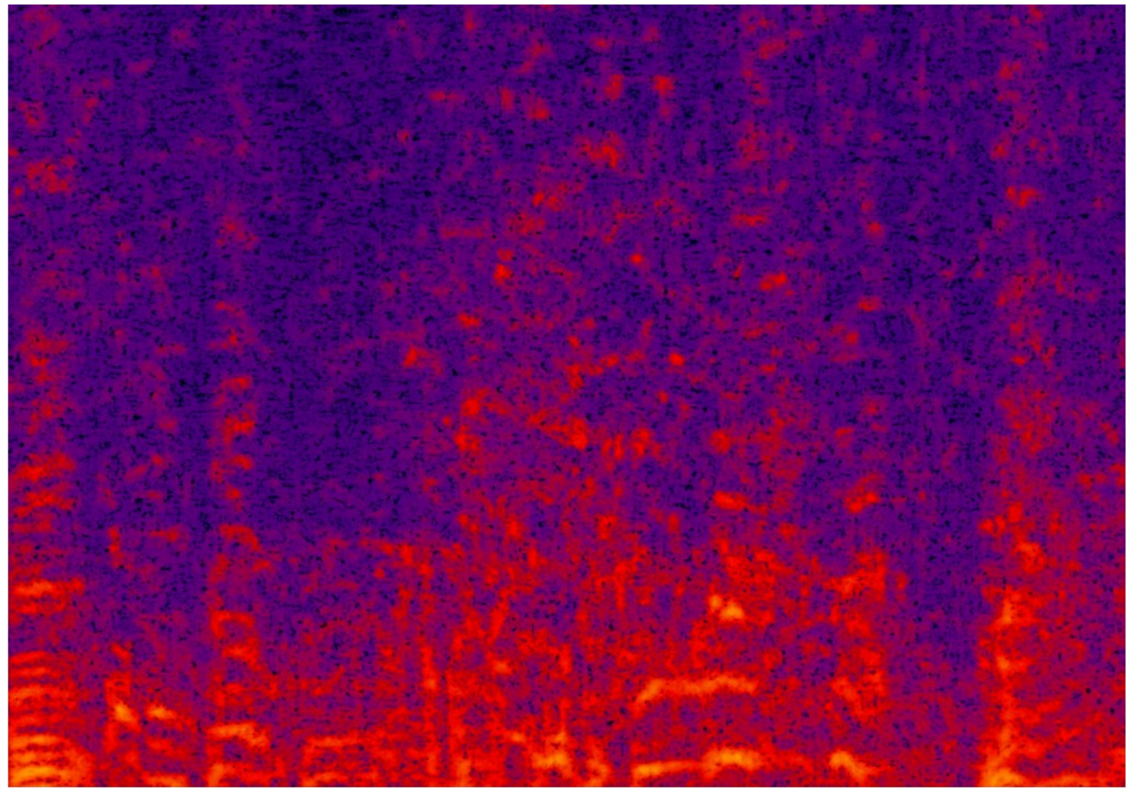
Metric	Best Emotion	Range (8 Emotions)	Interpretation
MCD ↓	Neutral (7.3 dB)	7.3−11.2	Measures spectral closeness Higher → more distortion
F0 RMSE ↓	Sleepy (13.8 Hz)	13.8−24.7	Measures pitch stability/accuracy
PESQ ↑	Neutral (2.3)	1.5−2.3	Perceptual quality (predicted)
MOSNet ↑	Neutral (2.9)	1.9−2.9	Predicted naturalness and Confirms feasibility
Key Trade-off	Neutral/Sleepy show high fidelity (low MCD).	Angry/Surprised show high pitch range (high F0 RMSE) but Whispering/Drunk show the lowest spectral fidelity (MCD≈11.2).	

Spectrograms of Prosody Transfer

Example Text: "Auch ein ungewolltes Kind ist ein wunderbares Geschenk."
(Even an unwanted child is a wonderful gift.)

- **Ground Truth (Angry):** High and variable F0 (pitch) contour. High energy peaks at key words (ungewolltes, wunderbares). Clear prosodic emphasis.
- **Synthesized (Angry):** Similar high F0 contour and timing. Energy distribution closely follows the ground truth. Minor spectral blur (token quantization noise).





Emotion-wise Analysis

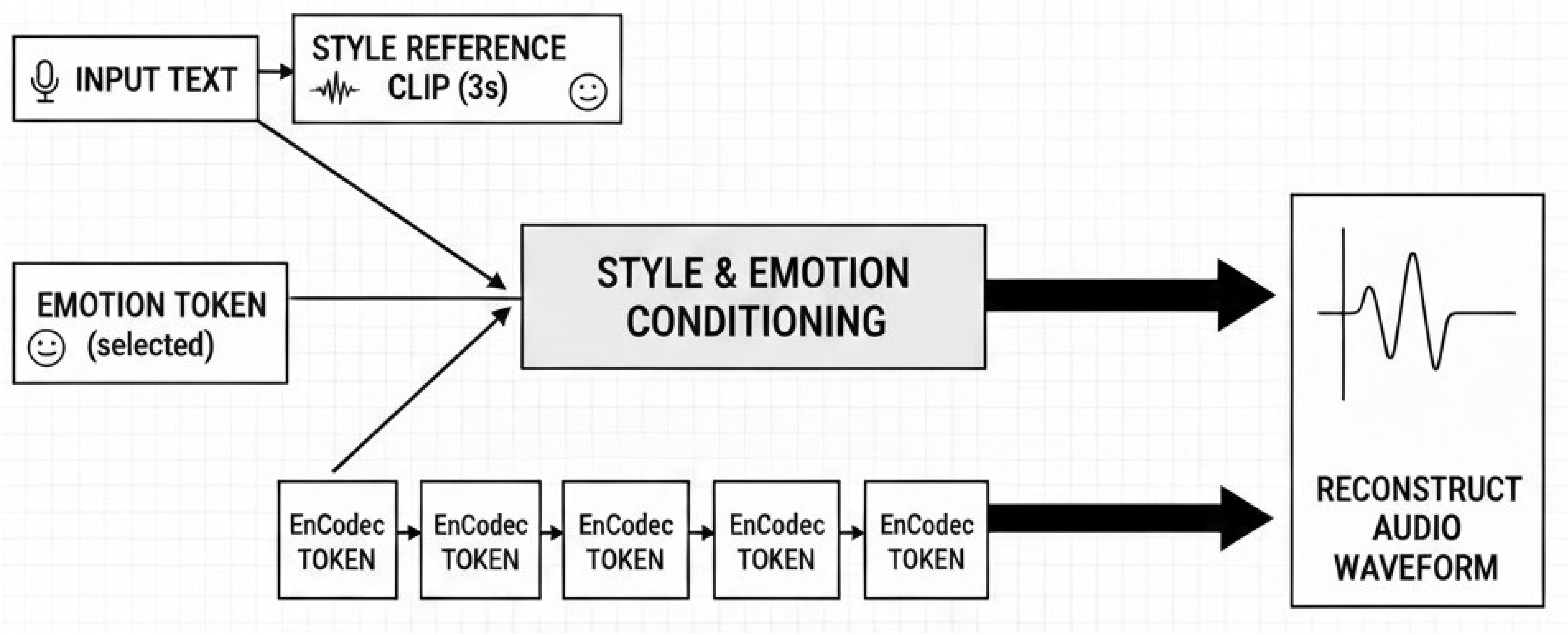
- Neutral & Sleepy → smooth prosody.
- Angry & Surprised → higher pitch variance.
- Drunk & Whisper → distorted spectrum.

Discussion

Finding	Interpretation	Impact
Emotion Transfer	The model learns and reproduces complex German prosody from limited data.	Shows that expressivity can be achieved with few-shot training.
Quality Gap	Synthesized speech sounds natural but still below human-level quality.	Indicates EnCodec tokens struggle to capture full emotional richness.
The VALL-E Edge	VALL-E effectively separates phoneme content from speaking style.	Confirms VALL-E’s potential for efficient emotional TTS in low-resource languages.

Inference Workflow

- Provide input text (optionally with a style reference)
- Extract the style embedding from a 3-second reference clip
- Select the desired emotion token
- Generate discrete EnCodec tokens → reconstruct the audio waveform



Conclusion and Future Directions

Conclusion

- **Dual-Controllable Architecture:** Developed the first emotion-controllable VALL-E-based model for German, using explicit Emotion Tokens and implicit Style Embeddings.

- **Feasibility for Low-Resource TTS:** Experimentally validated few-shot learning on the challenging ≈3-hour SLR110 dataset, proving that expressive speech is feasible with minimal data.

Future Work

- **Enhanced Regularization:** Implement data-free self-training or enhanced regularization to stabilize extreme style generation.
- **Contrastive Loss:** Implement a contrastive learning loss to force greater separation between the style and emotion latent spaces.
- **Cross-Lingual Transfer:** Investigate transfer learning capabilities to other low-resource languages like Dutch or Arabic using the same VALL-E backbone.

References

- Kim, Daegyeom, Hong, Seong-soo, and Choi, Yong-Hoon. ”SC VALL-E: Style-Controllable Zero-Shot Text to Speech Synthesizer.” ArXiv, 2023. Available: https://api.semanticscholar.org/CorpusID:259991058
- Barakat, H., Turk, O. & Demiroglu, C. Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. J AUDIO SPEECH MUSIC PROC. 2024, 11 (2024). https://doi.org/10.1186/s13636-024-00329-7
- S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, “Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers,” IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 705–718, 2025. doi: 10.1109/TASLPRO.2025.3530270.
- R. Liu, B. Sisman, and H. Li, “Reinforcement Learning for Emotional Text-to-Speech Synthesis with Improved Emotion Discriminability,” in Proceedings of Interspeech 2021, Aug. 2021, pp. 4648–4652. doi: 10.21437/Interspeech.2021-1236.
- T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, “Cross-Speaker Emotion Disentangling and Transfer for End-to-End Speech Synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1448–1460, 2022. doi: 10.1109/TASLP.2022.3164181.
- Wonbin Jung and Junhyeok Lee. E3-VITS: Emotional End-to-End TTS with Cross-speaker Style Transfer. Published online: 23 Jun 2023, last modified: 07 Jul 2023. https://openreview.net/forum?id=UZoYwJHZMzz.
- Skerry-Ryan, R.J., Stanton, D., Wu, Y., et al., “Tacotron: Towards End-to-End Speech Synthesis,” Interspeech, 2017, pp. 4006–4010, August. doi:10.21437/Interspeech.2017-1452.
- Lu, Hui, Wu, Zhiyong, Wu, Xixin, Li, Xu, Kang, Shiyin, Liu, Xunying, and Meng, Helen, “VAENAR-TTS: Variational Auto-Encoder Based Non-AutoRegressive Text-to-Speech Synthesis,” Interspeech, 2021, pp. 3775–3779, August. doi:10.21437/Interspeech.2021-2121.
- Karlapati, S., Karanasou, P., Lajszczak, M., et al., “CopyCat2: A single model for multi-speaker TTS and many-to-many fine-grained prosody transfer,” Amazon Science, 2022.
- Y. Korotkova, I. Kalinovskiy, and T. Vakhrusheva, “Word-level Text Markup for Prosody Control in Speech Synthesis,” in Proceedings of Interspeech 2024, Sept. 2024, pp. 2280–2284. doi: https://doi.org/10.21437/Interspeech.2024-71510.21437/Interspeech.2024-715.

Thank you

Mohammed Salah Al-Radhi

malradhi@tmit.bme.hu

Style-Controlled VALL-E for Few-Shot Emotional German TTS

EKÖP
Egyetemi Kutatói Ösztöndíj Program

ENFIELD