Objective Comparison of F5-TTS and VITS Trained on MARA and SWARA for Romanian Speech Synthesis

Oscar Gal
Telecommunications Department
Technical University of Cluj Napoca
Cluj-Napoca, Cluj
oscar.gal@campus.utcluj.ro

Mircea Giurgiu
Telecommunications Department
Technical University of Cluj Napoca
Cluj-Napoca, Cluj
mircea.giurgiu@com.utcluj.ro

***Abstract*—We present a controlled objective comparison of F5-TTS and VITS trained for Romanian synthesis with MARA and SWARA. The two corpora provide 69,421 paired utterances (77.91 h), including the now audio-complete 61,287-record SWARA manifest. Our contribution is a shared Romanian normalization/IPA front end and a file-level benchmark that evaluates both architectures with the same 15 texts, natural references, signal preparation, ASR, and seven metrics. The benchmark contains 330 synthesized–reference pairs: 15 per model under MARA conditioning and 150 per model over ten SWARA voices. F5-TTS has lower MCD, WER, and CER in both conditions. It leads four of seven measures on MARA and five of seven on SWARA; VITS retains stronger F0-contour correlation and the best same-speaker PESQ. F5-TTS is therefore the stronger overall system in this controlled matched-text benchmark.**

***Index Terms*—Romanian text-to-speech, F5-TTS, VITS, MARA, SWARA, objective evaluation**

I. Introduction

Text-to-speech (TTS) quality reflects the combined influence of data, conditioning, front ends, vocoders, and inference settings. VITS integrates a conditional variational autoencoder, flows, duration prediction, and adversarial waveform generation [1]. F5-TTS instead uses reference-conditioned optimal-transport flow matching with a Diffusion Transformer (DiT), implicit duration, and iterative generation [2]. Their architectural contrast provides a strong basis for a reproducible Romanian benchmark.

This paper contributes: (1) an audited account of the MARA and SWARA data; (2) one reproducible TTSTrainer front end shared by both training paths; and (3) a common-text, file-level evaluation of all available F5-TTS and VITS outputs using MCD, PESQ, STOI, WER, CER, F0-RMSE, and F0 correlation. The analysis is centered on generated speech and reports every metric distribution together with model-wise effect magnitudes. This yields a transparent ranking supported by spectral, perceptual, prosodic, and recognition-based evidence.

II. Architectures

A. VITS

VITS learns speech in a latent space. A posterior encoder maps a reference spectrogram to latent sequence z, while a text prior and normalizing flows model the same space. Monotonic alignment search links text and acoustic time. A stochastic duration predictor learns phoneme durations, and a HiFi-GAN-style decoder generates the waveform. Training combines reconstruction, Kullback–Leibler, duration, adversarial, and feature-matching losses; the posterior encoder and discriminator are omitted at inference [1]. The local configuration uses a multi-speaker embedding and an integrated waveform generator.

The separation between posterior and prior is central to the local training path. The posterior sees the complete acoustic observation and supplies latent samples, whereas the text prior must predict their distribution from IPA and speaker identity. The invertible flow makes these distributions compatible, and monotonic alignment search selects the highest-likelihood ordered path without external phoneme timings. The duration model is stochastic rather than a fixed length regressor, allowing multiple rhythms for one text. At inference, only the text prior, duration predictor, inverse flow, and decoder remain. The result is a compact single-stage synthesizer, but acoustic and waveform errors are coupled because the decoder is optimized adversarially with the rest of the network.

B. F5-TTS

Conditional flow matching learns a vector field that transports noise toward the data distribution. Matcha-TTS established efficient conditional-flow generation [3]; E2 TTS showed text-guided infilling without an explicit duration model [4]. F5-TTS refines this family with character embeddings, ConvNeXt V2 text refinement, a DiT conditioned by adaptive layer normalization, and Sway Sampling. During training, a random mel-spectrogram region is masked and reconstructed. During inference, the reference speech/transcript form the known region and the ODE solver generates the

TABLE I
Audited Romanian training data.

Corpus	Records	Spk.	Hours	Evidence
MARA	8,134	1	10.57	7,727 in train manifest
SWARA	61,287	41	67.34	Primary manifest complete
Total	69,421	42	77.91	Paired audio/text

target region [2]. F5-TTS predicts mel features and uses an external waveform vocoder, whereas VITS decodes its latent representation directly. Our measurements evaluate the finalized waveform from each complete inference path.

In more detail, the DiT estimates the velocity of an optimal-transport path between a noise sample and the target mel representation. The conditioning mel frames remain visible while the held-out span supplies the learning target; the text encoder provides the aligned symbolic context without a separate duration label. Numerical ODE integration follows the learned field at inference, and Sway Sampling redistributes solver steps toward more useful regions of that trajectory. Reference audio consequently controls voice and acoustic context directly. This is flexible for SWARA's many voices, although runtime depends on the number of solver steps and final waveform quality also depends on the separately trained vocoder.

III. Data and Preprocessing

A. MARA and SWARA

Table I reports the two corpora used here. MARA is a single-speaker Romanian audiobook corpus containing 8,134 paired 22.05-kHz, mono, 16-bit PCM utterances (10.57 h). Its curated training manifest contains 7,727 entries, giving a dense and acoustically consistent basis for single-speaker Romanian synthesis.

SWARA contains 61,287 primary manifest records (67.34 h) from 41 speaker codes, split into 49,030 training and 12,257 testing records. Every primary manifest record has a matching 22.05-kHz, mono, 16-bit PCM WAV. The many-speaker coverage complements MARA and enables the same Romanian linguistic front end to be tested under substantially broader voice conditioning.

B. TTSTrainer Text Front End

Both training paths use the supplied TTSTrainer pipeline. It applies Unicode NFC, maps Romanian cedilla characters to comma-below forms, standardizes white space and quotation marks, preserves prosodic punctuation, segments sentences, and expands dates, times, percentages, currencies, measurement units, cardinal numbers, decimals, abbreviations, and Romanian ordinal forms. Linguistic analysis uses the local Stanza Romanian model in the recorded environment. Pronunciation follows project overrides, then Romanian eSpeak IPA, then a flagged grapheme fallback. Stress is inherited from the eSpeak mark or a documented heuristic. F5 receives the IPA string with a character tokenizer; VITS sets `use_phonemes=false` because the manifest already contains IPA, avoiding a second phonemization pass.

Table II is a real TTSTrainer run rather than a hand-written example. It demonstrates that normalization happens before IPA conversion and that lexical stress remains explicit in the serialized sequence. The trace records four rules: abbreviation, unit, date, and time expansion.

C. Emotion, Focus, and Prosody Pipeline

Beyond pronunciation, TTSTrainer provides an expressive profile containing emotion, intensity, focus, prosody, interjection, and context layers. A local multilingual XLM-RoBERTa classifier first produces probabilities for eleven source emotions. These are aggregated into six project labels: joy, gratitude, and love become *happy*; anger, frustration, contempt, and disgust become *angry*; sadness, fear, surprise, and neutral retain their corresponding labels. Neutral is a positive prediction, while *unspecified* is an explicit abstention when the leading probability is below 0.40 or tied. A runner-up reaching at least half the leading probability is retained as a secondary label rather than discarded. This joint use of context-level emotion and locally focused words follows a current direction in context-aware emotive synthesis [5].

Continuous valence and arousal are distribution-weighted combinations of fixed per-emotion prototypes. Intensity starts from arousal and is only nudged by repeated exclamation/question marks and capitalized words, with a maximum boost of 0.15; scores below 0.45, from 0.45 to 0.65, and at least 0.65 map to low, medium, and high. Thus punctuation supports but cannot select an emotion. Sentence focus is computed separately from lexical stress. Explainable signals include capitalization, an intensifier's target, repetition, emotion-bearing words, the main predicate, and Romanian corrective constructions of the form "*nu X, ci Y*." Function words require stronger evidence, and explicit user focus overrides the rule score.

The prosody stage converts these annotations to model-independent controls. Arousal adjusts speaking rate, relative pitch, and energy within 0.80–1.20, 0.85–1.20, and 0.80–1.20, respectively; focused tokens receive an additional local pitch/energy boost. Punctuation supplies pauses from 150 ms for a comma to 600 ms for an ellipsis, dependency relations introduce clause-boundary pauses, and sentence type selects rising, falling, or continuation contours. Context smoothing stores a second annotation rather than overwriting the local prediction: it blends current and previous distributions with weights 0.7 and 0.3, skips predictions of confidence at least 0.75, resets at paragraph boundaries, and respects contrast/consequence markers. Interjection insertion is disabled by default. These expressive annotations are implemented and traceable. The present MARA/SWARA experiment establishes the neutral acoustic baseline, while the same annotations de-

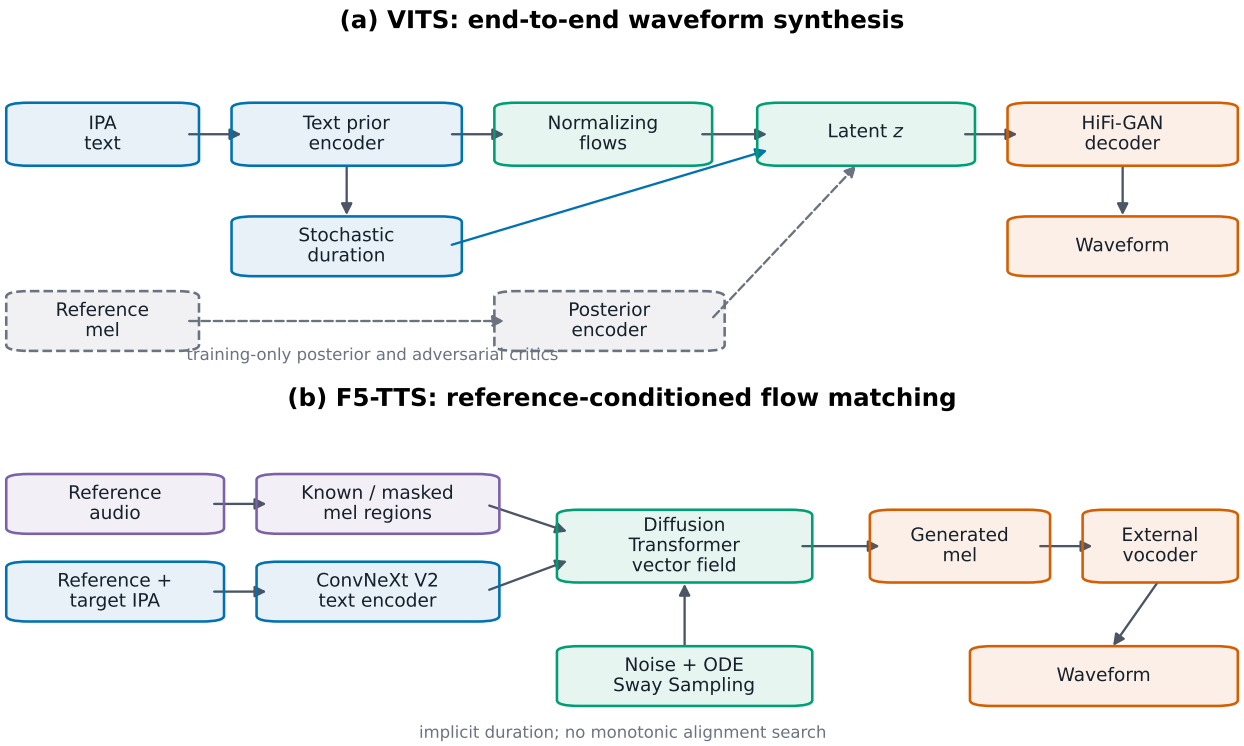


Fig. 1. Architecture-level comparison. Dashed VITS blocks are training-only; F5-TTS performs reference-conditioned mel infilling and requires an external vocoder. Both local systems receive the same TTSTrainer IPA representation.

TABLE II
Worked TTSTrainer preprocessing example.

Stage	Value
Raw	`Dr. Popescu a trimis 25 kg pe 12.07.2026, la ora 14:30.`
Normalized	Doctor Popescu a trimis douăzeci și cinci de kilograme pe doisprezece iulie două mii douăzeci și șase, la ora paisprezece și treizeci de minute.
IPA excerpt	`d' oktor pop' esku ' a trim' is ... k, ilogr' ame ... p' aisp... tr, eiz' etsh`
Manifest	`audio.wav|<accented IPA>|speaker`

fine a ready-made protocol for a balanced emotion-conditioned extension.

D. Audio Preparation

Both audited corpora are natively aligned at 22.05 kHz. The VITS feature configuration fixes a 22.05-kHz rate, 1024-sample FFT and window, 256-sample hop, 80 mel filters, and 0-Hz lower mel bound. The F5 preparation step preserves source WAVs through absolute symlinks, validates record integrity, applies the configured duration bound, and emits a two-column `audio_file|text` manifest for the official character-token preparation stage. F5 synthesis exports at 24 kHz; VITS exports at 22.05 kHz.

TTSTrainer also implements an optional deterministic realism chain: 70-Hz high-pass filtering, 30-dB endpoint trimming, 10% small-room convolution with 0.25-s decay, -55 dBFS room tone, -23 LUFS normalization, and a -1 dBFS peak ceiling. These are post-synthesis options, not training features. The objective benchmark instead evaluates the supplied, unmodified exports after the common metric preparation described in Section IV.

IV. Training and Evaluation Protocol

A. Training

The VITS log records an 86,405,228-parameter, four-GPU DDP run with FP16, seed 54321, initial generator/discriminator learning rates of 2×10^{-4}, 1,000 epochs, and 1,768,975 logged steps over about 8 d 22 h; the selected checkpoint is `best_model_79605.pth` at epoch 44. The F5 artifacts confirm successful Romanian fine-tuning and inference. Because the architectures optimize different objectives, the paper compares their common end product: frozen waveform exports evaluated through one identical file-level protocol.

B. Paired Test Construction

The curated common-text set contains 15 MARA prompts with complete natural, F5-TTS, and VITS coverage. We pair their natural recordings with 15 outputs per model under MARA conditioning and with 150 outputs per model from ten SWARA voice conditions. The resulting 330 pairs are matched by exact Unicode-decoded transcript, ensuring that every comparison uses identical linguistic content. MARA

conditioning measures same-speaker resemblance; SWARA conditioning broadens the test across ten requested voices while preserving the same target sentences. Together they provide complementary views of acoustic similarity and text intelligibility.

For metric computation, all signals are decoded to mono, resampled to 16 kHz, and endpoint-trimmed 40 dB below the utterance peak. There is no denoising, time stretching, or loudness normalization. Shorter signals are zero-padded for PESQ/STOI; their intrinsic timing penalties are retained.

C. Metrics

The evaluation follows the multidimensional practice supported by recent toolkits such as VERSA [6]. MCD uses 24 WORLD spectral-envelope mel-cepstral coefficients, excludes c_0, and applies DTW on 5-ms frames:

$$\mathrm{MCD} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(c_d^{(r)} - c_d^{(s)})^2}. \tag{1}$$

The aligned-frame distance is averaged per utterance [7]. WORLD DIO/StoneMask estimates F0 over 60–500 Hz. Jointly voiced frames on the same DTW path enter F0-RMSE and Pearson correlation.

Wideband PESQ measures reference/degraded perceptual quality [8]; STOI estimates intelligibility [9]. Both can penalize legitimate one-to-many timing/prosody and must not be interpreted as direct naturalness scores. WER and CER use one frozen Whisper-small Romanian decoder (beam 5, temperature 0, no previous-text conditioning) and one target normalizer. Natural references are transcribed by the same ASR to expose its floor. We show utterance distributions because legacy full-reference metrics may correlate weakly with human ratings for high-quality TTS [10], [11].

V. Experimental Results

Table III gives the means for every requested measure; no pair or metric is missing. Natural recordings produce WER 0.320 and CER 0.075 under the frozen ASR, confirming a non-zero recognition floor caused partly by literary names and punctuation normalization.

A. Quality and Intelligibility

Figure 2 shows all MCD, PESQ, and STOI observations. F5-TTS has lower MCD under MARA (6.755 vs. 6.870 dB) and SWARA (7.993 vs. 8.219 dB). Its ASR advantage is larger: WER/CER are 0.349/0.102 versus 0.557/0.160 on MARA and 0.242/0.067 versus 0.615/0.208 on SWARA. F5-TTS also has higher MARA STOI. Under SWARA, VITS has a small STOI advantage (0.151 vs. 0.139), while F5-TTS has higher PESQ (1.218 vs. 1.189). VITS obtains the best PESQ in the same-speaker MARA condition (1.210 vs. 1.059).

B. Prosody and Corpus Conditioning

On MARA, VITS tracks the reference contour better: F0-RMSE is 56.90 versus 61.74 Hz and correlation is 0.353 versus 0.099. On SWARA, F5-TTS has lower F0-RMSE (70.39 vs. 73.86 Hz), while VITS again has higher correlation (0.251 vs. 0.098). The MCD and F0 shift from MARA to SWARA reflects the intentional transition from the single MARA reference voice to diverse SWARA voice conditioning.

C. Effect Sizes and Distributional Reading

Relative to VITS, F5-TTS reduces mean MCD by 1.7% on MARA and 2.8% on SWARA. The recognition margins are substantially larger: F5-TTS reduces WER by 37.3% and CER by 36.0% on MARA, and by 60.7% and 67.6%, respectively, on SWARA. These reductions are the strongest evidence supporting the overall ranking because they occur under both conditioning regimes and both word- and character-level scoring. The natural-reference WER of 0.320 further provides a transparent ASR calibration point.

The opposing results are also material. On MARA, VITS improves F0-RMSE by 7.8% and raises F0 correlation from 0.099 to 0.353; its PESQ is 14.3% higher. On SWARA, VITS retains a 0.153 absolute correlation advantage and 8.9% higher STOI, whereas F5-TTS improves PESQ by 2.4% and F0-RMSE by 4.7%. Figure 2 also shows wide, overlapping utterance distributions and several high-error outliers. Thus a mean win is not equivalent to dominance on every sentence or speaker. Since the MARA and SWARA group sizes differ by a factor of ten, we compare models within each condition rather than pooling all 330 pairs into one headline mean.

The plots expose full utterance-level dispersion, and the conclusion relies on directional consistency plus effect magnitude. A future speaker-cluster bootstrap can complement these descriptive results with confidence intervals over the SWARA voice groups.

D. SWARA Voice-Level Comparison

Figure 4 shows that the aggregate intelligibility advantage is distributed across the requested voices. F5-TTS voice means occupy compact ranges of 0.189–0.314 WER and 0.045–0.111 CER; CAU gives the lowest errors. The VITS voices span 0.498–0.759 WER and 0.142–0.316 CER, with BVL giving its lowest errors. Thus the corpus-level result is supported by all ten available voice conditions per model rather than by one exceptionally favorable voice.

VI. Discussion

The radar profile in Fig. 3 summarizes, without averaging unlike units, the consistent division between systems. F5-TTS is more intelligible and spectrally closer to the available reference; VITS follows its F0 shape more closely and gives the best same-speaker PESQ. F5-TTS wins four of seven MARA dimensions and five of seven SWARA dimensions. Its largest practical margin is ASR intelligibility, especially on SWARA, whereas the VITS advantage is clearest in F0 correlation.

TABLE III
Objective means (15 MARA or 150 SWARA-conditioned pairs per system).

System/condition	MCD↓	PESQ↑	STOI↑	F0-RMSE↓	F0 corr.↑	WER↓	CER↓
F5-TTS / MARA	6.755	1.059	0.214	61.74	0.099	0.349	0.102
VITS / MARA	6.870	1.210	0.141	56.90	0.353	0.557	0.160
F5-TTS / SWARA	7.993	1.218	0.139	70.39	0.098	0.242	0.067
VITS / SWARA	8.219	1.189	0.151	73.86	0.251	0.615	0.208

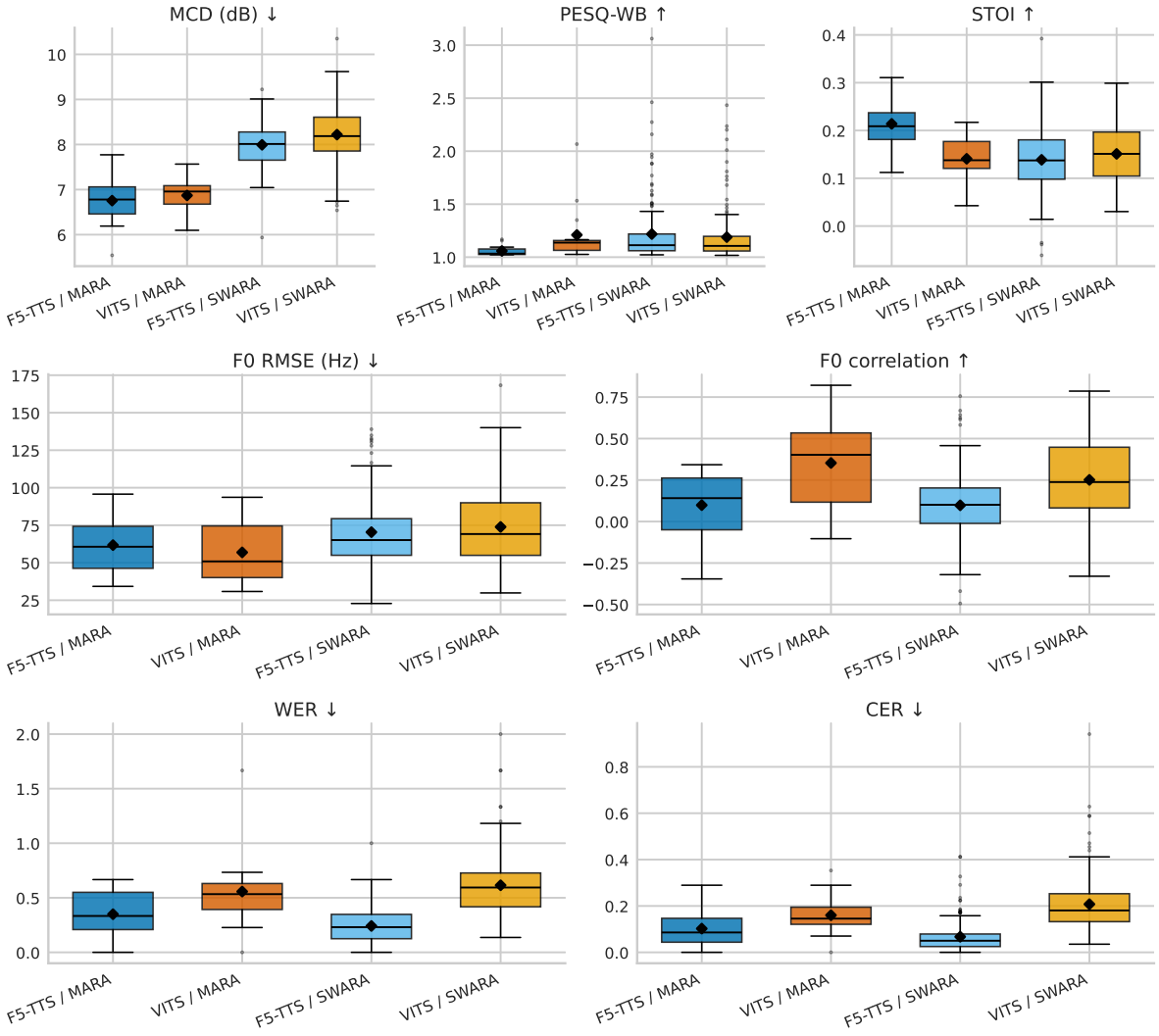


Fig. 2. Complete utterance-level objective results: MCD, PESQ, STOI, F0-RMSE, F0 correlation, WER, and CER. Diamonds are arithmetic means; box center lines are medians. F0 measures use jointly voiced, DTW-aligned frames; WER/CER use the frozen Romanian ASR.

The present evidence characterizes the complete delivered systems and reveals a consistent architecture-level performance profile. The next validation stage can deepen that result through listening judgments, speaker-similarity embeddings, paired uncertainty estimates, and an expanded common-text manifest. Those additions would connect the strong objective intelligibility gains to perceived naturalness, speaker fidelity, and preference.

The expressive pipeline creates a valuable second experimental axis. A controlled study can hold text and speaker constant while comparing predicted controls, manual emotion labels, and a neutral baseline across all six labels. WER/CER can verify content preservation; emotion agreement, valence/arousal error, pause duration, speaking rate, and focus-local F0/energy change can measure control realization. Blinded ratings of emotion, intensity, focus, naturalness, and speaker similarity would complete this objective profile.

The benchmark also defines a direct next stage: ex-

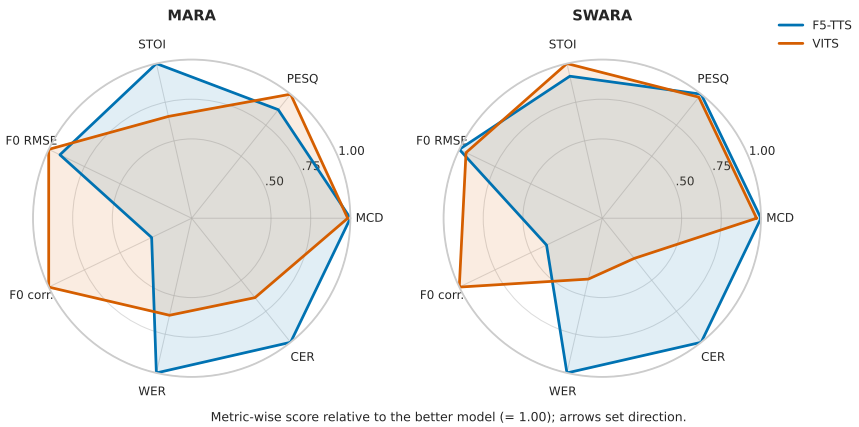


Fig. 3. Metric-wise relative performance. Within each corpus, the better system receives 1.00; the other is divided by the better value after orienting all axes so higher is better. This is a profile, not a composite quality score.

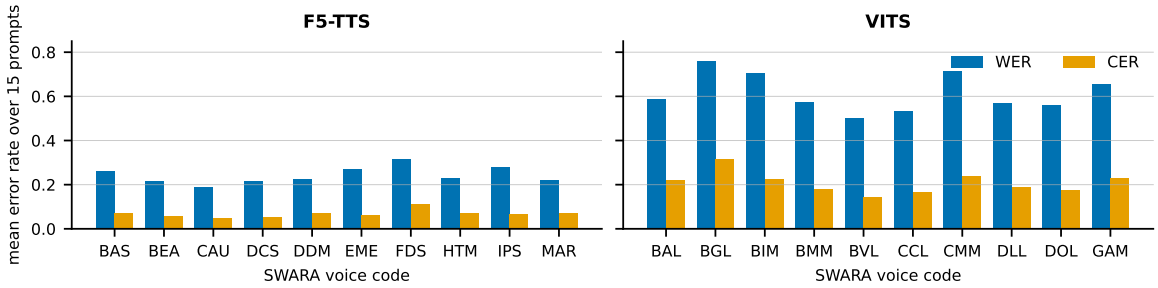


Fig. 4. SWARA voice-level intelligibility. Each bar is the mean over the same 15 prompts. The panels retain the ten available voice codes for each model, making both the level and consistency of WER/CER visible.

panded speaker-matched manifests, serialized inference settings, speaker-cluster confidence intervals, and randomized ratings of naturalness, intelligibility, similarity, and preference. This preserves the shared TTSTrainer front end while extending the objective evidence toward perception and deployment reproducibility.

VII. Conclusion

Using one preprocessing front end and 330 exact text-matched synthesis–reference pairs, F5-TTS is the better overall model for the present MARA/SWARA Romanian benchmark. It leads both conditions in MCD, WER, and CER and wins most of the seven requested dimensions. VITS remains preferable when reproducing the recorded F0 contour is primary and obtains the highest same-speaker PESQ. The shared TTSTrainer front end, audited corpora, complete metric distributions, and two conditioning regimes make this a reusable Romanian evaluation baseline. The next stage is an expanded speaker-matched set with paired confidence intervals and human listening judgments.

References

[1] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, vol. 139, 2021, pp. 5530–5540.

[2] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," in *Proc. ACL*, 2025, pp. 6255–6271.

[3] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-TTS: A fast TTS architecture with conditional flow matching," in *Proc. ICASSP*, 2024, pp. 11 341–11 345.

[4] S. E. Eskimez, X. Wang, M. Thakker *et al.*, "E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot TTS," in *Proc. IEEE SLT*, 2024, pp. 682–689.

[5] Y. Jeon, Y. Kim, J. Lee, and G. Lee, "Prompt-guided selective masking loss for context-aware emotive text-to-speech," in *Findings of NAACL*, 2025, pp. 638–650.

[6] J. Shi, H. jin Shim, J. Tian *et al.*, "VERSA: A versatile evaluation toolkit for speech, audio, and music," in *Proc. NAACL-HLT: System Demonstrations*, 2025, pp. 191–209.

[7] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Communications, Computers and Signal Processing*, vol. 1, 1993, pp. 125–128.

[8] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[10] C. Minixhofer, O. Klejch, and P. Bell, "TTSDS2: Robust objective evaluation for human-quality synthetic speech," in *Proc. 13th Speech Synthesis Workshop*, 2025, pp. 68–75.

[11] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "A review on subjective and objective evaluation of synthetic speech," *Acoustical Science and Technology*, vol. 45, no. 4, pp. 161–183, 2024.