

Voice Clone Similarity Report

Run #2 · upload · language: tr · generated locally

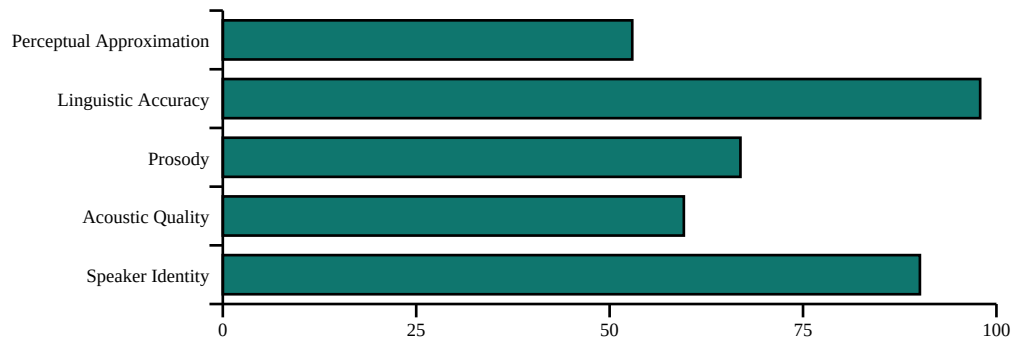
Voice closeness score

75.6%

The cloned voice is close to the original, with visible gaps in at least one metric domain.

This percentage is an engineering similarity score computed from available objective metrics. It is not a biometric-security bypass guarantee.

Domain similarity breakdown



Audio inputs

Audio	Duration	Sample rate
Original	71.57s	22050
Clone	61.01s	22050

AI engineering summary

Verdict:

The voice clone demonstrates a strong speaker identity match and high linguistic accuracy with good acoustic and perceptual quality, yet exhibits a notable duration mismatch and some spectral distortion.

Evidence:

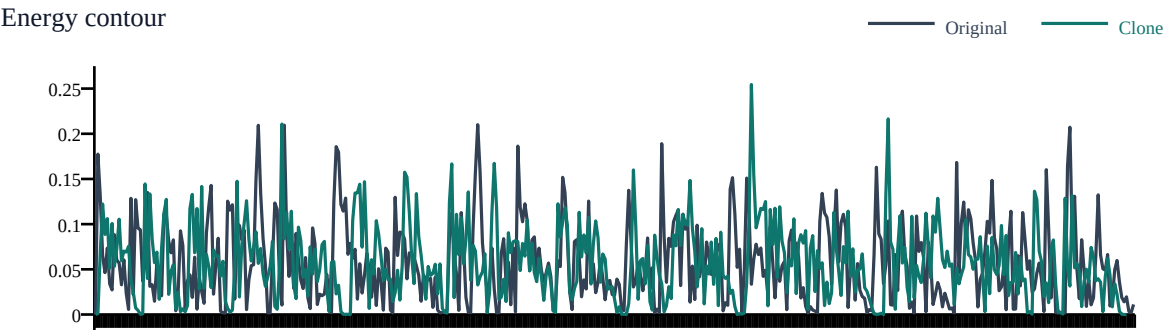
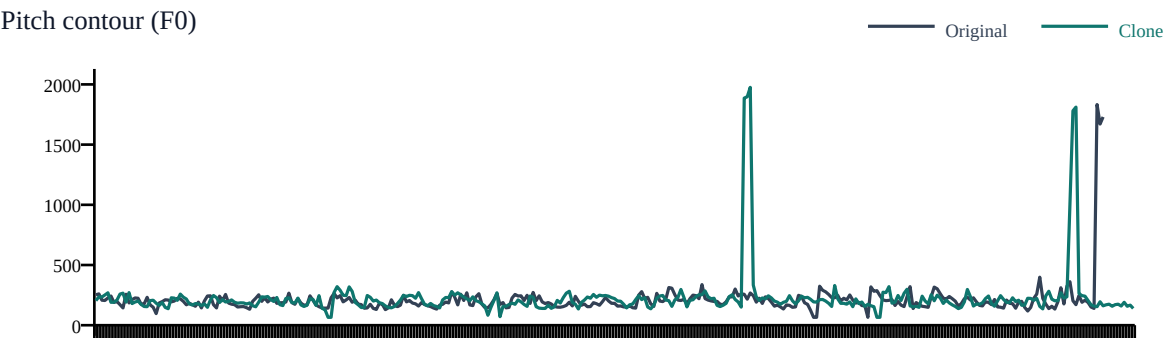
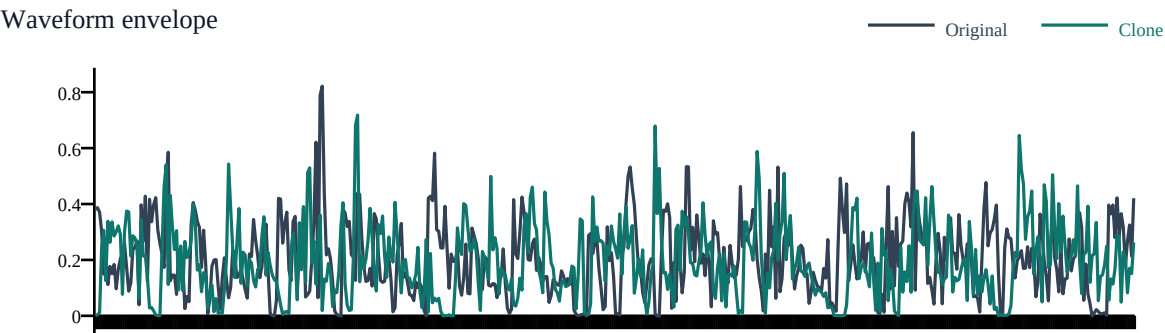
- Strongest match: ECAPA-TDNN speaker similarity is high (0.85), MFCC timbre similarity near perfect (0.997), and formant trajectory correlations exceed 0.83, supporting good voice identity preservation.
- Acoustic quality: Clone SNR is high (28.6 dB); however, Log-Spectral Distance is moderately elevated (20.9 dB) with a worst chunk reaching 59 dB, indicating spectral deviations.
- Prosody: Energy contour correlation is high (0.94), F0 contour correlation decent (0.72), but F0 RMSE is relatively large (136.6 Hz), and duration ratio is off at 0.85 with a 10.56 s error, reflecting temporal alignment issues.
- Linguistic accuracy is excellent, with word and character error rates below 3%.
- Perceptual metrics show strong NISQA overall MOS (~4.63) and moderate DNSMOS (~3.38), indicating good perceived quality.

Next action:

Focus on temporal alignment and duration modeling improvements to reduce speaking rate discrepancy and duration errors, and investigate spectral distortion sources to lower the log-spectral distance and worst chunk error for smoother acoustic quality.

Graph comparison

These plots mirror the dashboard views and are generated from the stored analysis series.

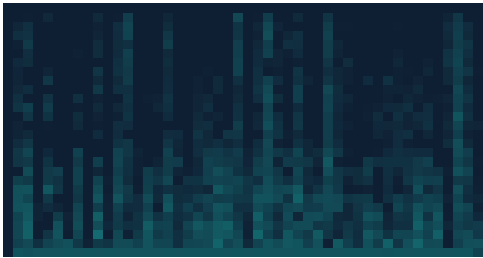


Per-chunk drift

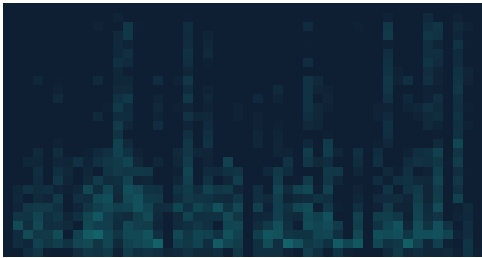


Spectrogram preview

Original



Clone



Metric evidence

Domain	Metric	Value	Direction	What it means
Acoustic Quality	Band Energy Ratio air_8000_plus	0.499 ratio	near_1	Compares high-frequency air and breath detail. Values near 1 preserve natural airiness.
Acoustic Quality	Band Energy Ratio low_0_300	0.727 ratio	near_1	Compares very low-frequency energy, including rumble and voice body. Values near 1 match the original.
Acoustic Quality	Band Energy Ratio presence_4000_8000	0.972 ratio	near_1	Compares presence and crispness frequencies. Values near 1 preserve articulation brightness.
Acoustic Quality	Band Energy Ratio speech_300_4000	1.136 ratio	near_1	Compares the main speech intelligibility band. Values near 1 preserve spoken clarity balance.
Acoustic Quality	Clone Clipping Rate	0.000 fraction	lower	Percentage of samples near digital clipping. Lower means fewer hard distortion artifacts.
Acoustic Quality	Clone SNR Estimate	28.62 dB	higher	Estimates the clone's signal-to-noise separation. Higher usually means less hiss or background noise.
Acoustic Quality	Clone Silence Ratio	0.144 fraction	neutral	Share of very quiet frames in the clone. Helps reveal excessive pauses or dropped audio.
Acoustic Quality	Clone Spectral Flatness	0.019	neutral	Estimates noise-like texture in the clone spectrum. Higher can indicate hiss or synthetic roughness.

Domain	Metric	Value	Direction	What it means
Acoustic Quality	Clone THD Estimate	0.997 ratio	lower	Approximate harmonic distortion on a voiced segment. Lower means fewer tonal distortion artifacts.
Acoustic Quality	Log-Spectral Distance	20.91 dB	lower	Compares original and clone frequency shapes on voiced aligned frames after LUFS normalization. Lower means fewer spectral distortions.
Acoustic Quality	Spectral Convergence	1.207	lower	Measures how closely the clone spectrogram envelope matches the original on voiced aligned frames. Lower is cleaner.
Acoustic Quality	Worst Chunk Log-Spectral Distance	59.03 dB	lower	Highest chunk-level spectral distance. Lower means no individual split degraded badly.
Linguistic Accuracy	Normalized Character Error Rate	0.019	lower	Punctuation-stripped, Unicode-folded character mismatch. Lower catches smaller pronunciation issues.
Linguistic Accuracy	Normalized Word Error Rate	0.026	lower	Punctuation-stripped, Unicode-folded ASR word mismatch against the transcript. Lower means clearer intelligibility.
Linguistic Accuracy	Phoneme Character Error Rate	0.018	lower	Phoneme-level character error from G2P/IPA transcription when available. Lower is closer pronunciation.
Perceptual Approximation	DNSMOS Background MOS	4.134 MOS	higher	DNSMOS estimate of background/noise quality. Higher means cleaner background.
Perceptual Approximation	DNSMOS Overall MOS	3.376 MOS	higher	Independent non-intrusive DNSMOS quality estimate. Higher means stronger predicted perceived quality.
Perceptual Approximation	DNSMOS Signal MOS	3.638 MOS	higher	DNSMOS estimate of speech-signal quality. Higher means fewer speech distortions.
Perceptual Approximation	NISQA Coloration	4.311 MOS	lower	Predicted muffling or spectral coloration from NISQA. Lower is better.
Perceptual Approximation	NISQA Discontinuity	4.619 MOS	lower	Predicted interruptions or unnatural breaks from NISQA. Lower is better.
Perceptual Approximation	NISQA Loudness	4.563 MOS	neutral	Predicted loudness quality from NISQA. Use with other MOS dimensions.
Perceptual Approximation	NISQA Noisiness	4.442 MOS	lower	Predicted noise impairment from NISQA. Lower is better.
Perceptual Approximation	NISQA Overall MOS	4.626 MOS	higher	Predicted human-like overall speech quality from a local NISQA model. Higher is better.
Prosody	Clone Pitch Range	1956.3 Hz	neutral	Pitch span of the clone recording. Compare with original to spot monotone or exaggerated delivery.
Prosody	Clone Speaking Rate	152.4 wpm	neutral	Estimated words per minute for the clone. Compare with original to spot pacing drift.
Prosody	Clone Voiced Ratio	0.796 fraction	neutral	Fraction of frames detected as voiced speech. Helps spot dropouts or unstable voicing.

Domain	Metric	Value	Direction	What it means
Prosody	Duration Error	10.56 s	lower	Absolute time difference between original and clone. Lower means better timing match.
Prosody	Duration Ratio	0.852 ratio	near_1	Clone duration divided by original duration. Near 1 means matching tempo.
Prosody	Energy Contour Correlation	0.940	higher	Compares loudness movement over time. Higher means closer stress and emphasis patterns.
Prosody	F0 Contour Correlation	0.725	higher	Compares pitch movement over time. Higher means the clone follows the original intonation better.
Prosody	F0 Distribution KL Divergence	0.185	lower	Compares overall pitch distribution shape. Lower means the clone uses a similar pitch range and density.
Prosody	F0 RMSE	136.6 Hz	lower	Average pitch error in Hertz after alignment. Lower means closer pitch matching.
Prosody	HNR Difference	1.367 dB	lower	Difference in harmonic-to-noise ratio between original and clone. Lower means closer voice periodicity and breath/noise balance.
Prosody	Jitter Difference	0.001 fraction	lower	Difference in cycle-to-cycle pitch perturbation. Lower means closer vocal stability.
Prosody	Original Pitch Range	1956.3 Hz	neutral	Pitch span of the original recording. Used as context for expressiveness.
Prosody	Original Speaking Rate	129.9 wpm	neutral	Estimated words per minute for the original. Context for pacing.
Prosody	Shimmer Difference	0.002 fraction	lower	Difference in cycle-to-cycle amplitude perturbation. Lower means closer amplitude stability.
Speaker Identity	ECAPA-TDNN Speaker Distance	0.151	lower	Inverse of ECAPA speaker similarity. Lower means closer speaker identity.
Speaker Identity	ECAPA-TDNN Speaker Similarity	0.849	higher	Biometric-grade speaker embedding cosine similarity from SpeechBrain ECAPA. Higher means closer speaker identity.
Speaker Identity	Formant F1 Trajectory Correlation	0.853	higher	Compares first-formant movement, related to vowel openness and vocal tract shape. Higher is closer.
Speaker Identity	Formant F2 Trajectory Correlation	0.872	higher	Compares second-formant movement, related to tongue position and vowel identity. Higher is closer.
Speaker Identity	Formant F3 Trajectory Correlation	0.851	higher	Compares third-formant movement, reflecting finer vocal tract resonance. Higher is closer.
Speaker Identity	Formant F4 Trajectory Correlation	0.832	higher	Compares fourth-formant movement, reflecting upper resonance detail. Higher is closer.
Speaker Identity	MFCC Timbre Distance	0.003	lower	Inverse of MFCC timbre similarity. Lower means closer vocal color.

Domain	Metric	Value	Direction	What it means
Speaker Identity	MFCC Timbre Similarity	0.997	higher	Lightweight timbre similarity from MFCC statistics. Higher means closer vocal color.
Speaker Identity	Timbre Consistency Index	0.988	higher	Measures whether timbre similarity stays stable across chunks. Higher means less vocal drift.

Recommended publication wording: “Our clone scored 75.6% similarity to the original on our objective voice evaluation suite.”