

PhonePaint — Localized Phoneme Editing

Tom Brenner, PhD, physicalchemistrytom@gmail.com, +491714777620

Summary

PhonePaint is a speech-processing pipeline that performs **targeted phoneme correction** in unconstrained recordings.

Operating on arbitrary speech, PhonePaint uses off-the-shelf transcription and alignment tools. The novelty lies in selective phoneme editing: a small (<9M) masked model performs inpainting in Encodec space. No voice cloning or end-to-end conversion with heavy generative models.

- Identity preservation: waveform copied outside corrected (selected) phones
- Exact alignment, frame-synchronous with original audio

Current use case: correction of lateralized alveolar fricatives (s, z) & compensatory errors: pharyngeal production of postalveolar fricatives/affricates (sh, jh, ch) and velar stops (k, g).

Listen

Zero-shot phone correction audio samples (opens in browser):

<https://storage.googleapis.com/phone-paint-demo/demo.html>

Results utilizing **16 kHz VoiceCraft EnCodec** (s, z, ch, sh, jh & k examples provided) are given. Comparisons with VoiceCraft (https://jasonppy.github.io/VoiceCraft_web/) and Seed-VC (<https://github.com/Plachtaa/seed-vc>) are included.

Technical note (brief)

Reconstruction runs in the latent space of a neural audio codec (EnCodec); no separate vocoder; no voice cloning, full-utterance generation or conversion. A masked acoustic model fills in only the target phoneme frames, conditioned on a phone token. Surrounding frames are left unchanged.

Contact

If this capability is relevant to your team's work, I would appreciate feedback and am happy to discuss further.