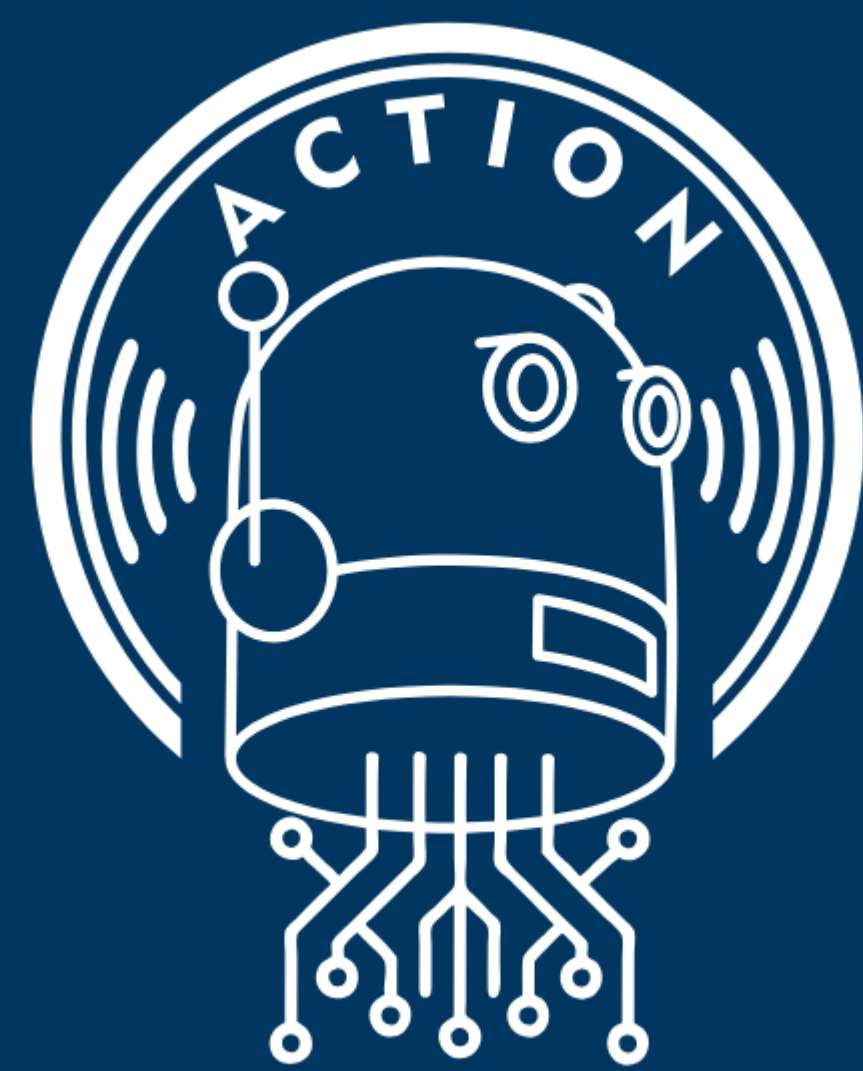




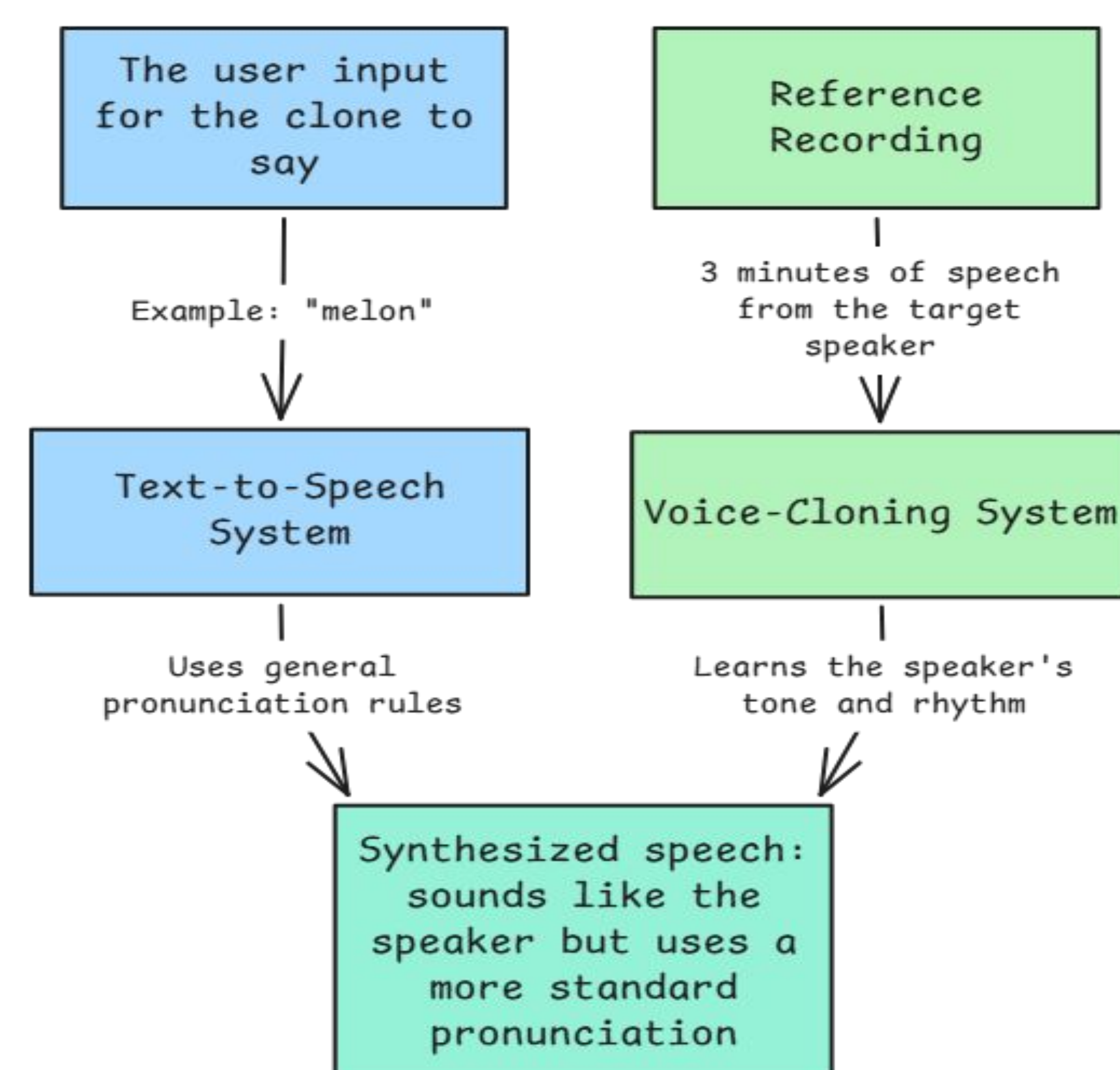
Speaker-Specific Pronunciation Cues for Text-to-Speech Voice-Clone Detection

Juan Nunez, Yigitcan Kaya, Giovanni Vigna, Christopher Kruegel
UC Santa Barbara

Motivation

- Public recordings and AI can be used to create realistic deepfakes of people using voice cloning technology.
- General deepfake detectors often fail on new voices, cloning tools, recording devices, or acoustic conditions. [1]
- Initial tests found that no phrase was consistently difficult for clones to synthesize – consistent with how **cloning systems separate the words being spoken from the voice saying them.**

Question: Can **speaker-specific pronunciations resist voice cloning** in Text-to-Speech synthesis?

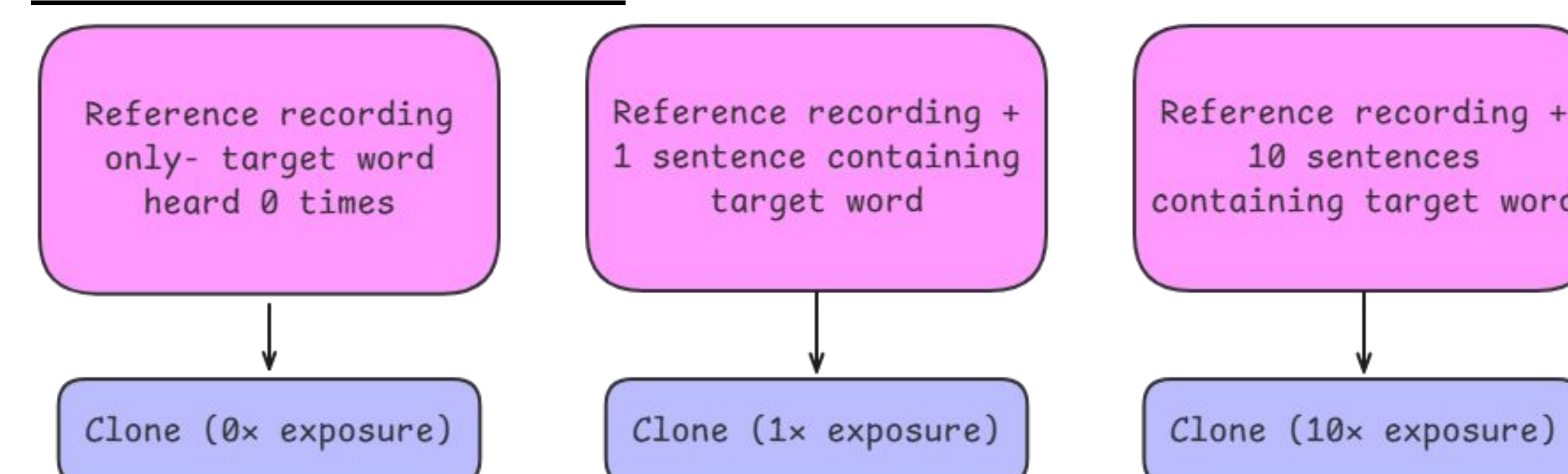


References:
[1] Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., & Böttinger, K. (2022). Does Audio Deepfake Detection Generalize?
[2] Zhao, G., Sonasat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2018). L2-ARCTIC: A Non-Native English Speech Corpus.

Methodology

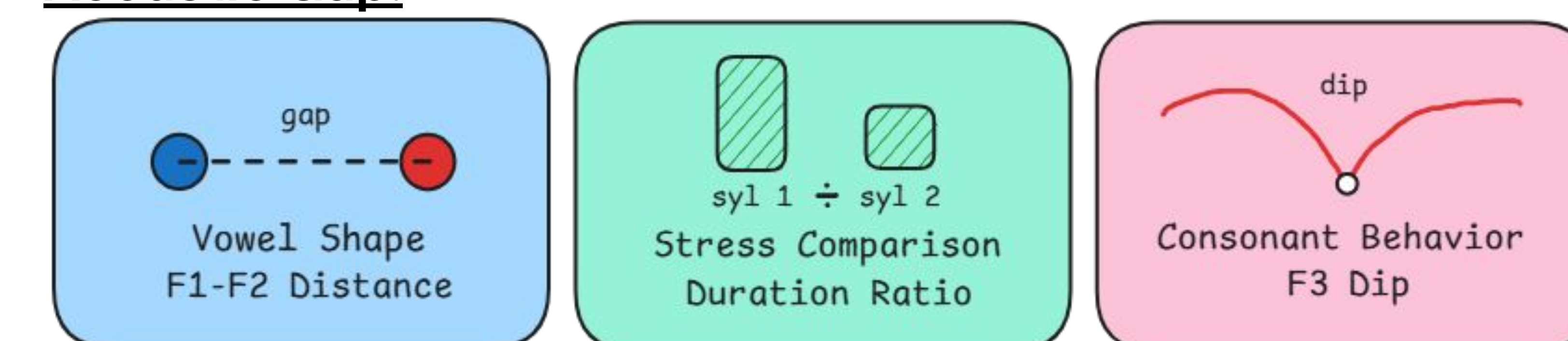
- Reference audio (3 minutes long) recorded from a member of the research team represented public data and was cloned using ElevenLabs and Cartesia.

Pronunciation Transfer:



- Each clone generated the same test sentences containing the target word.
- A blind rater judged whether the clone preserved the speaker's usual pronunciation.

Acoustic Gap:



3 different metrics used for calculating the gap between the genuine speaker and cloned recordings. This allowed us to capture how a person's habitual pronunciation can differ:

- Vowel quality: difference in vowel position (of the mouth)

$$\Delta(F1, F2) = \sqrt{(F1_{clone} - F1_{genuine})^2 + (F2_{clone} - F2_{genuine})^2}$$

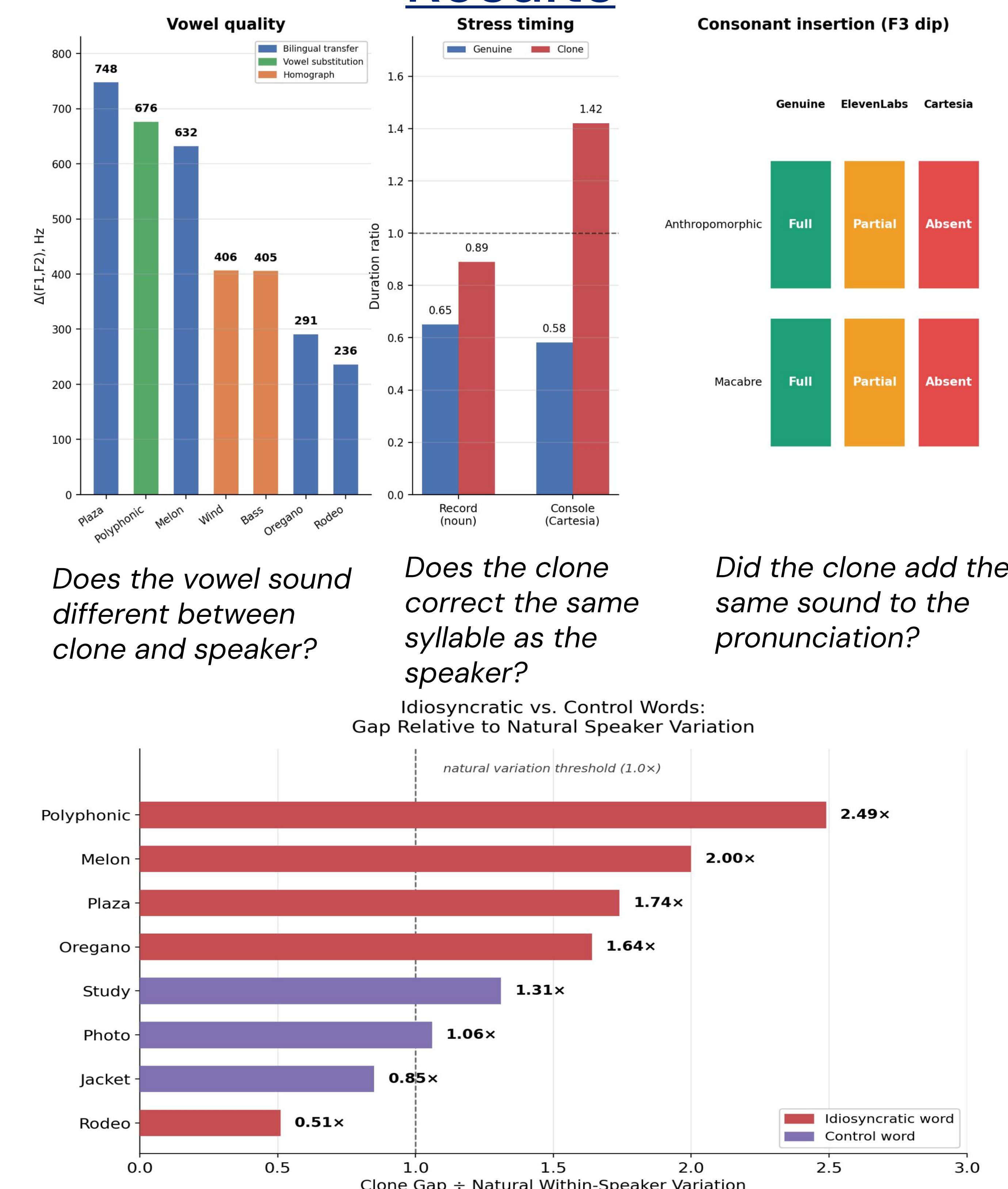
- Stress timing: relative duration of syllables in a word

$$R = \frac{T_{syllable 1}}{T_{syllable 2}} \quad (T = \text{duration of syllable})$$

- Consonant insertion: whether the expected "r" cue was fully present, partial, or absent

Consonant insertion was informed by L2-Arctic [2]

Results



Does the vowel sound different between clone and speaker?

Does the clone correct the same syllable as the speaker?

Did the clone add the same sound to the pronunciation?

Is the gap between the clone and speaker bigger than the natural variation between the target words?

Blind Rater Note: the listener, unaware of the exposure levels, saw no consistent increase on mispronunciation between dose levels – supporting the finding that more exposure to a speaker doesn't transfer their pronunciations.

Conclusion:

- Pilot Result:** voice clones often defaulted to standard English pronunciation for all words.
- Implication:** speaker-specific pronunciation may offer cues for personalized authentication.
- Future Work:** test with more speakers, language backgrounds, and cloning systems.