

Speak As Me: How Close a Generated Voice Gets to a Person, from One Recording

Madhu

Kansas City · Independent

ABSTRACT

We measured how closely a generated voice matches a person when it is built from a single recording of them, as part of an assistant that answers in a person's own words and voice while they are away. Likeness is placed on a scale that runs from a stranger to the person's own other recordings, and is multiplied by the share of words spoken correctly. From a clean recording the generated voice reached the level of the real speaker, with fewer than one word in a hundred wrong, generated faster than real time on the person's own machine. From a short, degraded recording of the kind people actually have, every approach tried fell to about four fifths of the way, and about a third of the remaining gap was then closed at more than twice the generating time. This brief states the results in relative terms and gives their limits. The measured figures are in the full results paper, shared on request. The system is proprietary: nothing here says which voice systems were used or how a recording is prepared.

1 The question

A stand-in that answers for a person needs to sound like them. A recording holds the sentences a person said; a voice can say sentences they never said. The question was how close a voice generated from one recording comes to the real one, both when the recording is clean and when it is the short, noisy clip a real person would supply.

2 How it was judged

Likeness is measured on a scale with two fixed ends for each speaker: the similarity of other speakers to this one at the bottom, and the similarity of the speaker's own held-out recordings at the top. The words are checked by transcribing the generated speech. The two are multiplied into one score, and reported under two conditions: a clean sample, and a shorter sample degraded like a phone or video recording. In both, the comparison is with the real, clean voice. A separate measure of audio quality must not fall, and generation must stay quick enough for conversation. The speakers are a small set of public-domain recordings.

3 What was found

Table 1: Results in relative terms.

Measure	Result
Likeness from a clean recording	at the level of the real speaker
Words spoken wrongly	fewer than one in a hundred
Speed of generation	faster than real time
Likeness from a degraded recording	about four fifths of the way to the speaker
Measured audio quality, degraded against clean	about a quarter lower
Remaining gap closed on degraded recordings	about a third
Generating time for that gain	more than twice as long

With a clean recording the problem is close to solved for this measure. With a degraded one, the approaches tried were within a few points of one another, and the noise in the sample came through in the generated voice. The recording given to the system mattered more than the choice of system.

4 Limits

The likeness measure hears the colour of a voice and not its rhythm or emphasis; it is necessary and not sufficient, and no listening test with people has been run. The test set is a handful of speakers and a short list of sentences in one language, so a difference of a point or two is noise. The degraded condition is simulated. No result has been reproduced by a third party.

5 Use

The system is intended for a voice the user owns or has explicit permission to use. The stand-in states that it is an AI when sincerely asked.

6 The figures

The scores for every configuration under both conditions, with the size of the test, are in the full results paper. It is shared on request to the author.