We sincerely thank all the reviewers for their helpful comments. We attempt to address their concerns below.

Additional evaluations and human judgment

• All reviewers recommend additional evaluations. • Reviewer 1 also suggests human evaluator opinion tests. • Reviewer 2 asks if humans are able to visualize faces for voices. • Reviewers 2 and 3 wonder if the reconstructed faces do indeed capture the speaker's ID. • Reviewer 3 recommends comparison to baselines. While we are unable to compare to existing baselines, since there is no prior work on the topic, we have run additional experiments which we hope address the rest of the reviewers' concerns under this topic. These results will be included in the paper and the experimental setups and data made public.

a. Do the generated faces actually capture the speaker's ID: The GAN is, in fact, optimized for face identification – a critical component of the loss is provided by a face-ID system (which is separately trained, and not optimized with the GAN). To explicitly address the question, we ran the face-ID system on a set of faces derived from novel speech recordings that are not part of the training set. The test is a "closed-set" test, in the sense that the IDs are expected to be one of the 924 IDs represented in the ID system. Table 1 shows the top-1 and top-5 identification results. While not perfect, the results clearly show that the faces generated from the voices clearly do capture the ID of the speaker.

Table 1: Results of closed-set recognition

Total number of samples	# of samples per class	# of classes	Top-1 accuracy	Top-5 accuracy
4620	5	924	61.7%	82.3%

b. Human judgment test: Studies in experimental psychology have shown that humans *are* indeed significantly better than chance at matching voices to faces, e.g. [1].

We ran a human judgment test, as suggested by the reviewer. A total of 20 volunteers (who had no prior knowledge of this work) were each presented 20 trials. Each trial comprised a voice and two GAN faces, one of which was produced from the voice. The subject was required to identify the generated face for the voice. To eliminate extraneous factors, we tried to match covariates, e.g. both faces were always the same gender. Trial voices and faces were not repeated for any subject. Table 2 shows the results. The results are much better than chance, with a P value below 0.001, and are similar to performances reported in human studies, e.g. [1].

Table 2: Human subject accuracy in matching voices to generated faces

Minimum accuracy	Maximum accuracy	Mean accuracy	Standard deviation
55.0%	75.0%	67.5%	8.1

[1] Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. Journal of Experimental Psychology: Human Perception and Performance, 39(2), 307-312

Why GAN architectures?

We attempted a variety of regression and GAN architectures. The GAN architecture that included a Face-ID loss turned out to be the most effective. The key component was the face-ID system that provides a loss to optimize the generator.

Predicting age from voice, which was not shown with high accuracy (reviewer 2)

Age prediction from voice is a well-studied problem, with mean absolute errors of less than 5 years being achievable by systems trained with age-labeled voice data.

In our setup, the challenge was the lack of appropriate training data, as the reviewer rightly points out. Voxceleb, has many voice recordings and face images for each person, however there is no information about the correspondence in age between them. Thus, when we pair a voice with a face during training, the voice may have been recorded when the subject was young, while the facial image may have been taken when they were much older. As a result, there really is no assurance of learning age associations, although remarkably, some association seems to be learned, presumably due to some degree of correspondence in the training data.

What would be an application for such a task? (reviewer 2)

We have had expressions of interest from the entertainment and gaming industries (e.g. to generate avatars). The research is directly funded by law enforcement agencies in the US, although they only expect to use some carefully-vetted aspects of it. The research itself was not, however, targeted at a specific application, but to determine if visual information beyond the obvious covariates could somehow be extracted from voice. The answer seems to be affirmative.