Iterative Text-based Editing of Talking-heads Using Neural Retargeting: Supplementary Material

XINWEI YAO, Stanford University
OHAD FRIED, Stanford University
KAYVON FATAHALIAN, Stanford University
MANEESH AGRAWALA, Stanford University

	Group 1	Group 2	Mean Difference	Adjusted p-value
Short Phrase	Fried [2019] (> 1 hr)	Fried [2019] (< 5 min)	-0.3127	**0.0065**
	Fried [2019] (> 1 hr)	Modified Fried (> 1 hr)	0.319	**0.0051**
	Fried [2019] (> 1 hr)	Ground-truth	0.6798	**0.001**
	Fried [2019] (> 1 hr)	Ours (< 5 min)	0.3067	**0.0072**
	Fried [2019] (< 5 min)	Modified Fried (> 1 hr)	0.6317	**0.001**
	Fried [2019] (< 5 min)	Ground-truth	0.9926	**0.001**
	Fried [2019] (< 5 min)	Ours (< 5 min)	0.6194	**0.001**
	Modified Fried (> 1 hr)	Ground-truth	0.3609	**0.0011**
	Modified Fried (> 1 hr)	Ours (< 5 min)	-0.0123	0.9
	Ground-truth	Ours (< 5 min)	-0.3732	**0.001**
Full Sentence	Fried [2019] (> 1 hr)	Fried [2019] (< 5 min)	-0.0185	0.9
	Fried [2019] (> 1 hr)	Modified Fried (> 1 hr)	0.3303	**0.001**
	Fried [2019] (> 1 hr)	Ground-truth	1.2914	**0.001**
	Fried [2019] (> 1 hr)	Ours (< 5 min)	0.6224	**0.001**
	Fried [2019] (< 5 min)	Modified Fried (> 1 hr)	0.3488	**0.001**
	Fried [2019] (< 5 min)	Ground-truth	1.3098	**0.001**
	Fried [2019] (< 5 min)	Ours (< 5 min)	0.6408	**0.001**
	Modified Fried (> 1 hr)	Ground-truth	0.961	**0.001**
	Modified Fried (> 1 hr)	Ours (< 5 min)	0.292	**0.0058**
	Ground-truth	Ours (< 5 min)	-0.669	**0.001**

Table 1. Adjusted p-values for all pair-wise comparisons in user studies. "Mean Difference" is mean score of Group 2 minus that of Group 1. Significant pair-wise comparisons (p<=0.05) have bolded p-values.

	Likert response (%)						
Condition	5	4	3	2	1	Mean	'Real'
Baseline	3.5	12.3	7.3	30.1	46.8	2.0	15.8%
Ours (automatic)	4.1	19.2	16.4	31.5	28.8	2.4	23.3%
Ours (refined)	5.8	27.0	13.3	30.3	23.6	2.6	32.8%

Table 2. User study results for videos with synthesized audio. We compare a video retiming baseline with our method with and without manual refinement. We report percentage of each answer on a 5-Point Likert scale, as well as mean score and percent of video that received a score of 4 or 5 ('real'). Both our pipelines outperform the baseline algorithm, which suggests that synthesizing video to match audio increases realism, even when using synthesized speech audio.

ACM Reference Format:
Xinwei Yao, Ohad Fried, Kayvon Fatahalian, and Maneesh Agrawala. 2020. Iterative Text-based Editing of Talking-heads Using Neural Retargeting: Supplementary Material. 1, 1 (November 2020), 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ADJUSTED P-VALUES FOR USER STUDIES

We apply Tukey's range test for pair-wise comparisons, and report the p-values of each pair (adjusted for multiple testing) in Table 1.

2 SYNTHETIC VOICE USER STUDY

Using the same design as our other user studies, our synthetic voice user study aims to rule out the hypothesis that using a synthesized voice in our tool is already so unrealistic, that there is no point in employing our sophisticated video synthesis technique. We recruited 73 participants to view 8 videos each, and compare a baseline method that linearly retimes the target video to match the length of the speech given by the edit, to results from an early version of our tool both automatically generated and with manual refinement (Table 2). The difference between conditions is statistically significant (Kruskal-Wallis test, $p < 10^{-9}$). Both our automatic synthesis pipeline and our pipeline with manual refinement perform better than the baseline algorithm (Tukey's range test, $p = 0.02$ and $p = 0.001$ respectively). These results suggest that synthesizing video to match audio increases realism, even when using synthesized speech audio.

REFERENCES

Authors' addresses: Xinwei Yao, xinwei.yao@cs.stanford.edu, Stanford University, Department of Computer Science; Ohad Fried, ohad@cs.stanford.edu, Stanford University, Department of Computer Science; Kayvon Fatahalian, kayvonf@cs.stanford.edu, Stanford University, Department of Computer Science; Maneesh Agrawala, maneesh@cs.stanford.edu, Stanford University, Department of Computer Science.

XXXX-XXXX/2020/11-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn