On Learning Associations of Faces and Voices

Supplementary Material

Changil Kim[1], Hijung Valentina Shin[2], Tae-Hyun Oh[1], Alexandre Kaspar[1], Mohamed Elgharib[3], and Wojciech Matusik[1]

[1] MIT CSAIL
[2] Adobe Research
[3] QCRI

Abstract. This document supplements the main paper and contains additional information therein referred.

Keywords: face-voice association · multi-modal representation learning

A.1 Data Collection for Human Performance

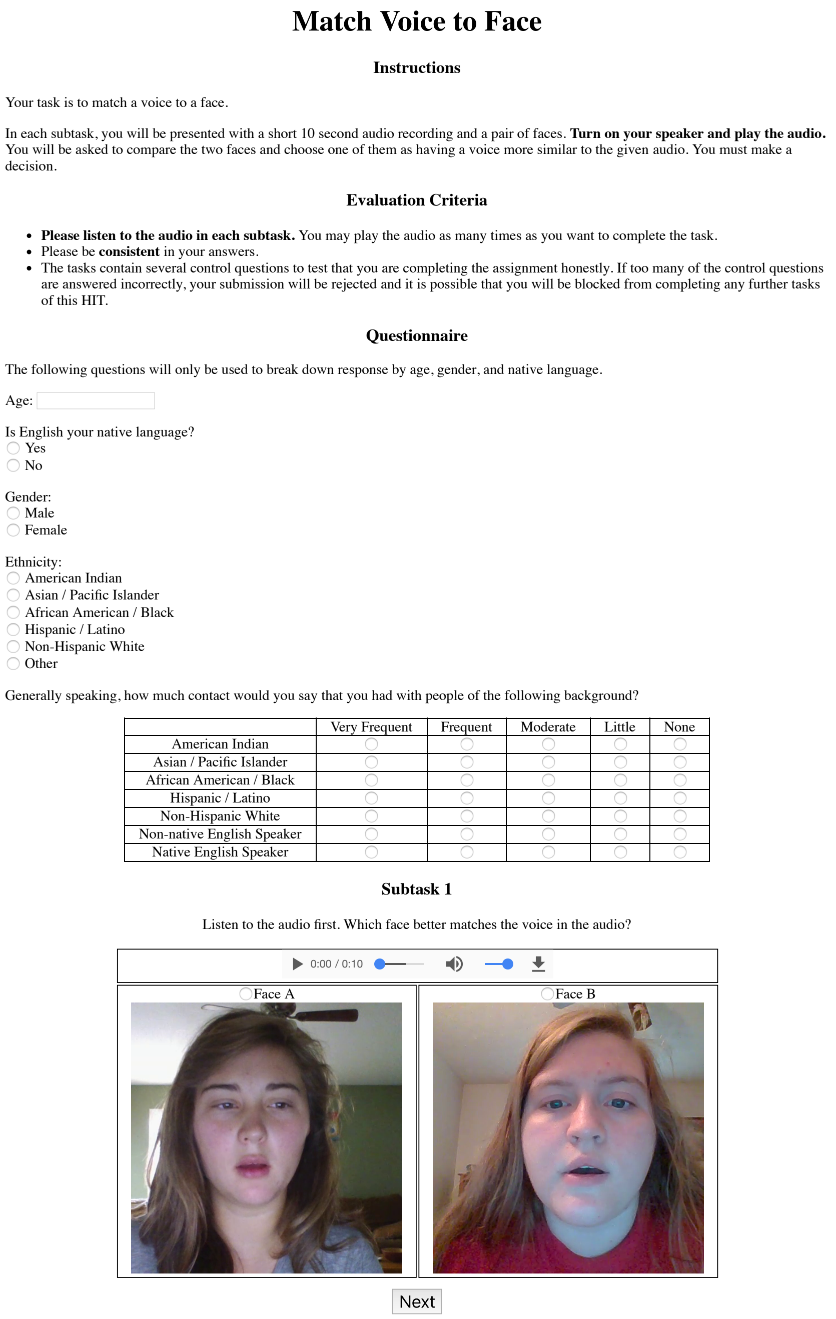
Both the data acquisition and user study were carried out through web applications deployed via Amazon Mechanical Turk. In the following, we present further details of the two tasks.

A.1.1 User Study

Figs. A.1 and A.2 show the questionnaire and example subtasks we used for Experiments 1–3 and Experiment 4, respectively, in Section 3 of the main paper. Actual subtasks are randomized every run.

A.1.2 Dataset Acquisition

Fig. A.3 shows the instructions for data collection. Every participant was requested to read the instructions carefully and to consent to the use of the collected dataset for research purposes. Fig. A.4 shows the questionnaire for demographic information and an example recording session. In order to encourage constant reading speed, words are sequentially highlighted in the script, similar to popular karaoke interfaces. Furthermore, to normalize the head position, we provide facial markers where participants can align their face to a centered front-facing position. Feedback about the alignment is provided using the `clmtrackr` library, a JavaScript implementation of the face tracking model of Saragih et al. [4]. Participants can repeat the recording session until they are satisfied. From the collected video recordings of them speaking, we extract still face images and ten-second-long audio clips containing their voices. We manually cleaned the collected data, for example, removing recordings with loud background noise or low audio/video quality. A few example face images are shown in Fig. A.5; for audio playback, browse our dataset at http://facevoice.csail.mit.edu. The text is chosen from the following pool:



Match Voice to Face

Instructions

Your task is to match a voice to a face.

In each subtask, you will be presented with a short 10 second audio recording and a pair of faces. **Turn on your speaker and play the audio.** You will be asked to compare the two faces and choose one of them as having a voice more similar to the given audio. You must make a decision.

Evaluation Criteria

- **Please listen to the audio in each subtask.** You may play the audio as many times as you want to complete the task.
- Please be **consistent** in your answers.
- The tasks contain several control questions to test that you are completing the assignment honestly. If too many of the control questions are answered incorrectly, your submission will be rejected and it is possible that you will be blocked from completing any further tasks of this HIT.

Questionnaire

The following questions will only be used to break down response by age, gender, and native language.

Age:

Is English your native language?
○ Yes
○ No

Gender:
○ Male
○ Female

Ethnicity:
○ American Indian
○ Asian / Pacific Islander
○ African American / Black
○ Hispanic / Latino
○ Non-Hispanic White
○ Other

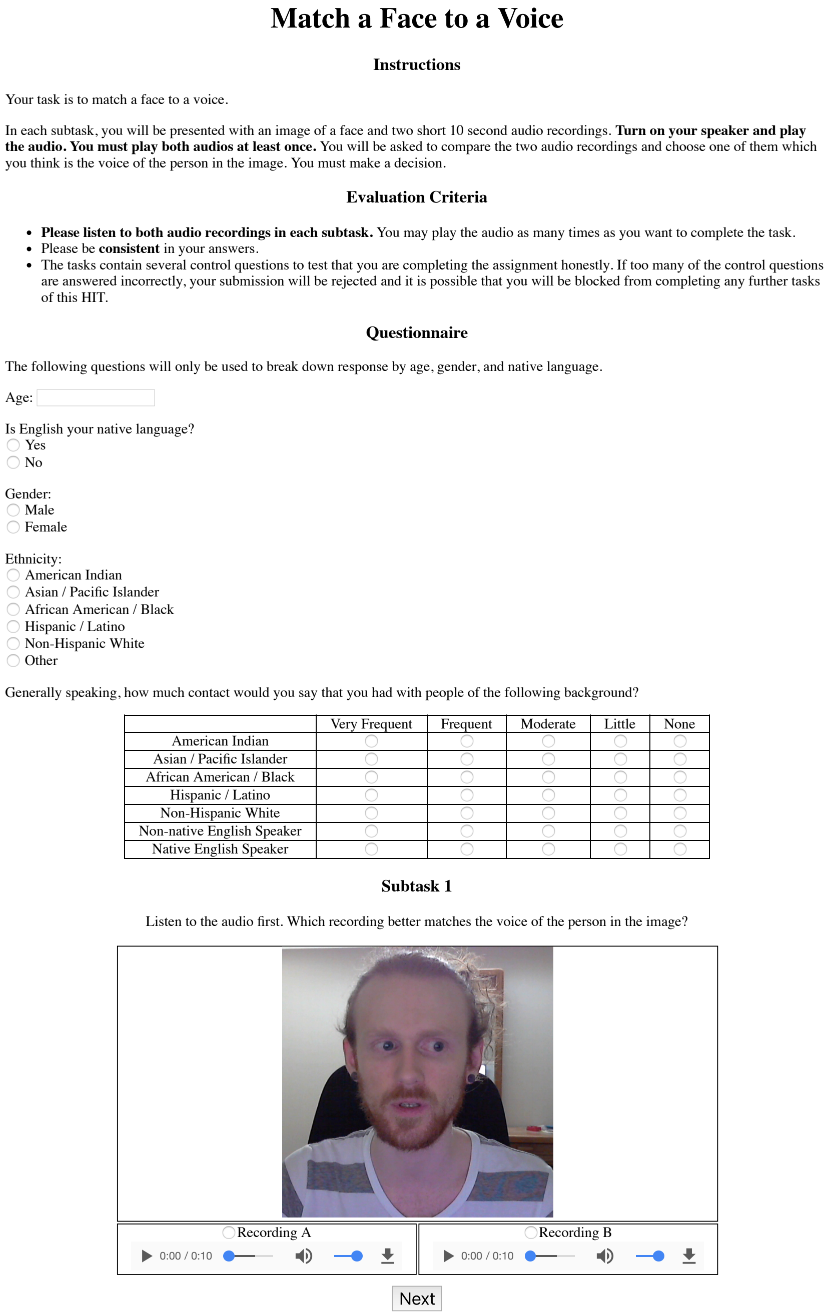
Generally speaking, how much contact would you say that you had with people of the following background?

	Very Frequent	Frequent	Moderate	Little	None
American Indian	○	○	○	○	○
Asian / Pacific Islander	○	○	○	○	○
African American / Black	○	○	○	○	○
Hispanic / Latino	○	○	○	○	○
Non-Hispanic White	○	○	○	○	○
Non-native English Speaker	○	○	○	○	○
Native English Speaker	○	○	○	○	○

Subtask 1

Listen to the audio first. Which face better matches the voice in the audio?

Fig. A.1. Screenshot of our user study questionnaire used for Experiments 1–3 (V → F). The answers were collected through the web interface of Amazon Mechanical Turk.



Match a Face to a Voice

Instructions

Your task is to match a face to a voice.

In each subtask, you will be presented with an image of a face and two short 10 second audio recordings. **Turn on your speaker and play the audio. You must play both audios at least once.** You will be asked to compare the two audio recordings and choose one of them which you think is the voice of the person in the image. You must make a decision.

Evaluation Criteria

- **Please listen to both audio recordings in each subtask.** You may play the audio as many times as you want to complete the task.
- Please be **consistent** in your answers.
- The tasks contain several control questions to test that you are completing the assignment honestly. If too many of the control questions are answered incorrectly, your submission will be rejected and it is possible that you will be blocked from completing any further tasks of this HIT.

Questionnaire

The following questions will only be used to break down response by age, gender, and native language.

Age:

Is English your native language?
○ Yes
○ No

Gender:
○ Male
○ Female

Ethnicity:
○ American Indian
○ Asian / Pacific Islander
○ African American / Black
○ Hispanic / Latino
○ Non-Hispanic White
○ Other

Generally speaking, how much contact would you say that you had with people of the following background?

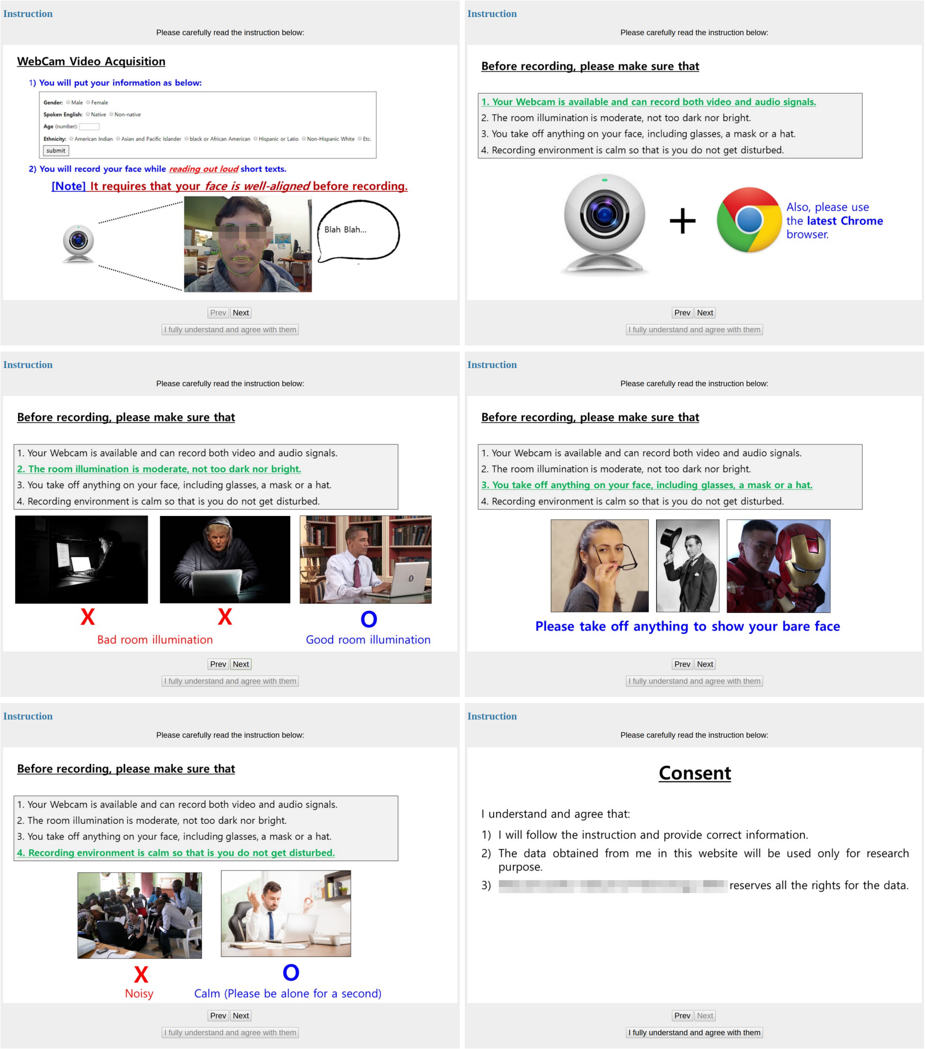
	Very Frequent	Frequent	Moderate	Little	None
American Indian	○	○	○	○	○
Asian / Pacific Islander	○	○	○	○	○
African American / Black	○	○	○	○	○
Hispanic / Latino	○	○	○	○	○
Non-Hispanic White	○	○	○	○	○
Non-native English Speaker	○	○	○	○	○
Native English Speaker	○	○	○	○	○

Subtask 1

Listen to the audio first. Which recording better matches the voice of the person in the image?

○ Recording A　○ Recording B

Fig. A.2. Screenshot of our user study questionnaire used for Experiments 4 (F → V). The answers were collected through the web interface of Amazon Mechanical Turk.



Instruction

Please carefully read the instruction below:

WebCam Video Acquisition

1) You will put your information as below:

Gender: ○ Male ○ Female

Spoken English: ○ Native ○ Non-native

Age (number):

Ethnicity: ○ American Indian ○ Asian and Pacific Islander ○ black or African American ○ Hispanic or Latio ○ Non-Hispanic White ○ Etc.

2) You will record your face while *reading out loud* short texts.

[Note] It requires that your *face is well-aligned* before recording.

Instruction

Please carefully read the instruction below:

Before recording, please make sure that

1. Your Webcam is available and can record both video and audio signals.
2. The room illumination is moderate, not too dark nor bright.
3. You take off anything on your face, including glasses, a mask or a hat.
4. Recording environment is calm so that is you do not get disturbed.

Also, please use the **latest Chrome** browser.

Instruction

Please carefully read the instruction below:

Before recording, please make sure that

1. Your Webcam is available and can record both video and audio signals.
2. The room illumination is moderate, not too dark nor bright.
3. You take off anything on your face, including glasses, a mask or a hat.
4. Recording environment is calm so that is you do not get disturbed.

Instruction

Please carefully read the instruction below:

Before recording, please make sure that

1. Your Webcam is available and can record both video and audio signals.
2. The room illumination is moderate, not too dark nor bright.
3. You take off anything on your face, including glasses, a mask or a hat.
4. Recording environment is calm so that is you do not get disturbed.

Please take off anything to show your bare face

Instruction

Please carefully read the instruction below:

Before recording, please make sure that

1. Your Webcam is available and can record both video and audio signals.
2. The room illumination is moderate, not too dark nor bright.
3. You take off anything on your face, including glasses, a mask or a hat.
4. Recording environment is calm so that is you do not get disturbed.

X
O
Noisy
Calm (Please be alone for a second)
Prev
Next
I fully understand and agree with them

Instruction

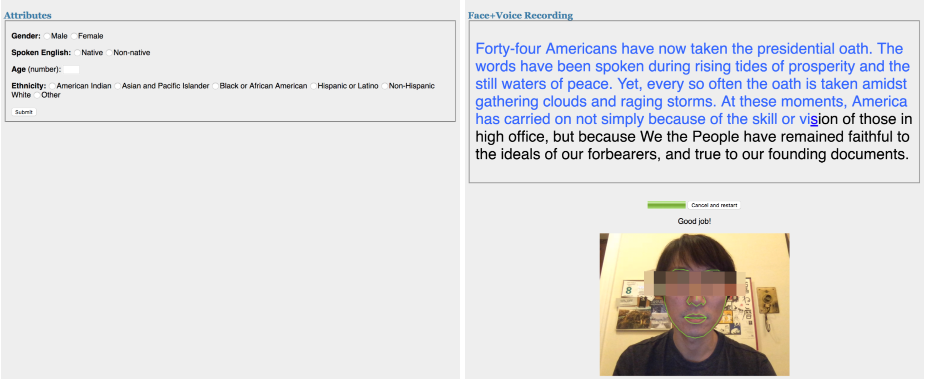
Please carefully read the instruction below:

Consent

I understand and agree that:

1) I will follow the instruction and provide correct information.
2) The data obtained from me in this website will be used only for research purpose.
3) [redacted] reserves all the rights for the data.

Fig. A.3. Screenshots of the instructions used for collecting our dataset for user studies. The web application was deployed through Amazon Mechanical Turk. Part of screenshots are masked out for anonymity.



Attributes

Gender: ○ Male ○ Female

Spoken English: ○ Native ○ Non-native

Age (number): □

Ethnicity: ○ American Indian ○ Asian and Pacific Islander ○ Black or African American ○ Hispanic or Latino ○ Non-Hispanic White ○ Other

Face+Voice Recording

Forty-four Americans have now taken the presidential oath. The words have been spoken during rising tides of prosperity and the still waters of peace. Yet, every so often the oath is taken amidst gathering clouds and raging storms. At these moments, America has carried on not simply because of the skill or vision of those in high office, but because We the People have remained faithful to the ideals of our forbearers, and true to our founding documents.

Good job!

Fig. A.4. Screenshots of a recording session of our dataset for user studies. The web application was deployed through Amazon Mechanical Turk. Part of screenshots are masked out for anonymity.

- "Forty-four Americans have now taken the presidential oath. The words have been spoken during rising tides of prosperity and the still waters of peace. Yet, every so often the oath is taken amidst gathering clouds and raging storms. At these moments, America has carried on not simply because of the skill or vision of those in high office, but because We the People have remained faithful to the ideals of our forbearers, and true to our founding documents."
- "That we are in the midst of crisis is now well understood. Our nation is at war, against a far-reaching network of violence and hatred. Our economy is badly weakened, a consequence of greed and irresponsibility on the part of some, but also our collective failure to make hard choices and prepare the nation for a new age. Homes have been lost; jobs shed; businesses shuttered. Our health care is too costly; our schools fail too many; and each day brings further evidence that the ways we use energy strengthen our adversaries and threaten our planet."
- "In reaffirming the greatness of our nation, we understand that greatness is never a given. It must be earned. Our journey has never been one of short-cuts or settling for less. It has not been the path for the faint-hearted - for those who prefer leisure over work, or seek only the pleasures of riches and fame. Rather, it has been the path for the risk-takers, the doers, the makers of things - some celebrated but more often men and women obscure in their labor, who have carried us up the long, rugged path towards prosperity and freedom."
- "For us, they fought and died, in places like Concord and Gettysburg; Normandy and Khe Sahn. Time and again these men and women struggled and sacrificed and worked till their hands were raw so that we might live a better life. They saw America as bigger than the sum of our individual ambitions; greater than all the differences of birth or wealth or faction."

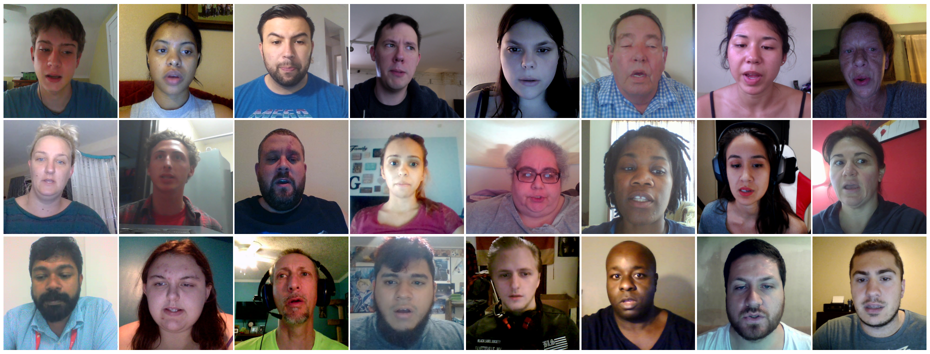


Fig. A.5. A few face samples from our collected dataset. See the accompanied video for voice playback.

– "To the Muslim world, we seek a new way forward, based on mutual interest and mutual respect. To those leaders around the globe who seek to sow conflict, or blame their society's ills on the West - know that your people will judge you on what you can build, not what you destroy. To those who cling to power through corruption and deceit and the silencing of dissent, know that you are on the wrong side of history; but that we will extend a hand if you are willing to unclench your fist."
– "For as much as government can do and must do, it is ultimately the faith and determination of the American people upon which this nation relies. It is the kindness to take in a stranger when the levees break, the selflessness of workers who would rather cut their hours than see a friend lose their job which sees us through our darkest hours. It is the firefighter's courage to storm a stairway filled with smoke, but also a parent's willingness to nurture a child, that finally decides our fate."
– "So let us mark this day with remembrance, of who we are and how far we have traveled. In the year of America's birth, in the coldest of months, a small band of patriots huddled by dying campfires on the shores of an icy river. The capital was abandoned. The enemy was advancing. The snow was stained with blood. At a moment when the outcome of our revolution was most in doubt, the father of our nation ordered these words be read to the people:"
– "America. In the face of our common dangers, in this winter of our hardship, let us remember these timeless words. With hope and virtue, let us brave once more the icy currents, and endure what storms may come. Let it be said by our children's children that when we were tested we refused to let this journey end, that we did not turn back nor did we falter; and with eyes fixed on the horizon and God's grace upon us, we carried forth that great gift of freedom and delivered it safely to future generations."

A.2 Evaluations on Machine Performance

In this section, further evaluations and visualizations of our learned representations omitted from Section 4.5 of the main paper are provided. We conclude this section with additional discussions.

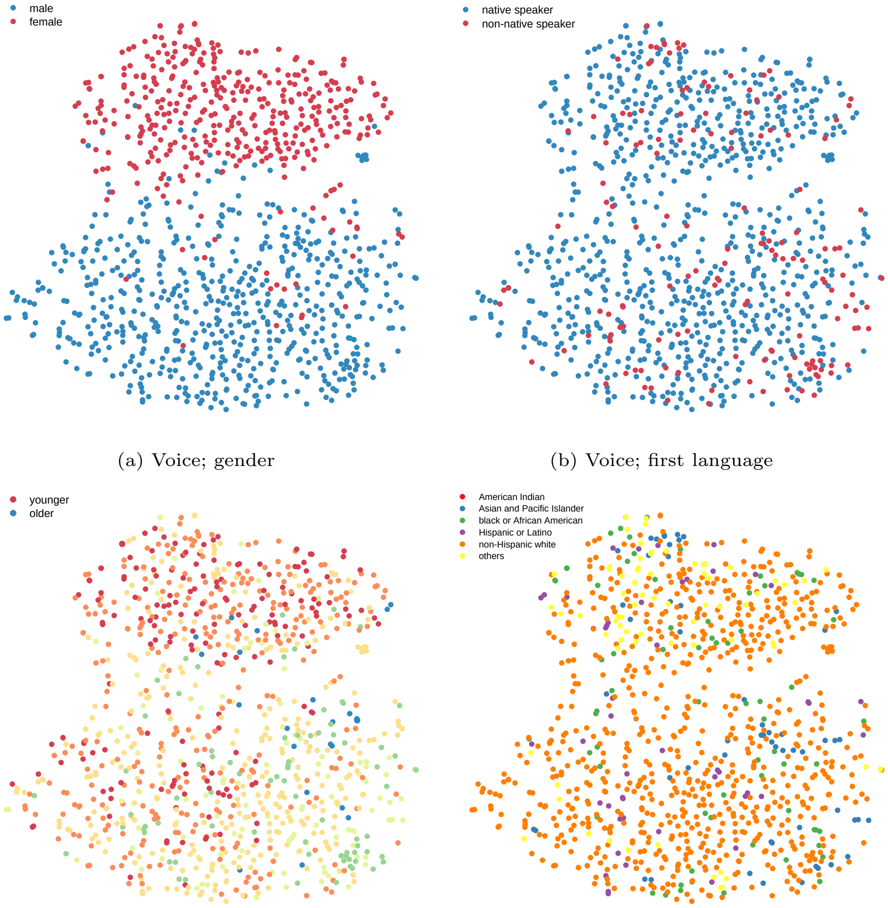
A.2.1 Further Evaluations on the Learned Representation

The t-SNE visualizations. Figs. A.6 and A.7 show the t-SNE visualization [2] of our learned voice and face representation, respectively. We drew 1,000 random samples and used our annotations to color-code the sample points according to their four demographic attributes. See Fig. 1 of the main paper for the t-SNE visualized with face/voice identities. Note that our network has not seen any of the demographic attributes during training.

As discussed in the main paper, the learned representation forms the clearest clusters regarding gender (Figs. A.6a and A.7a), which explains the performance drop when the samples are constrained by gender. While correlated with gender, age shows a distinct grouping from gender (Figs. A.6c and A.7c). In particular in face representations, Fig. A.7c shows the age distributed orthogonal to gender: it increases from bottom to top while the gender is split horizontally. The t-SNE visualization does not reveal such strong clustering regarding the first language or the ethnic group (Figs. A.6bd and A.7bd), and presents only small clusters scattered across the projection. As also noted in the main paper, such absence of clustering does not rule out the existence of additional information encoded in our learned representations, which we argue with additional evidences in the following.

More evaluations of linear classifiers on our representations. Table A.1 summarizes the quality of linear classifiers for demographic attributes on our learned representations, similar to Table 4 of the main paper, to demonstrate what information our representation encodes.

As evidenced by a high classification precision, the representation provides the most distinctive information for gender classification, which is consistent with the distributions observed in Figs. A.6a and A.7a. Age is a continuous attribute as demonstrated in Figs. A.6a and A.7a, and grouping into a discrete set of ranges (as in Table 4) makes the classification results more conservative: i.e., Fig. A.7f shows overall smooth transition in age, but far from perfect ordering especially in mid-ages, resulting in less decisive age classification results shown in Table A.1. We note that the analyses of Table A.1 and Figs. A.6 and A.7 (as well as Table 4 and Fig. 1 in the main paper) are complementary to, and consistent with, each other. t-SNE is an unsupervised method for visualization which typically reveals dominant information encoded in the representation, while the experiment in Table A.1 (and Table 4 of the main paper) exploits supervised information to reveal hidden information in the representation.



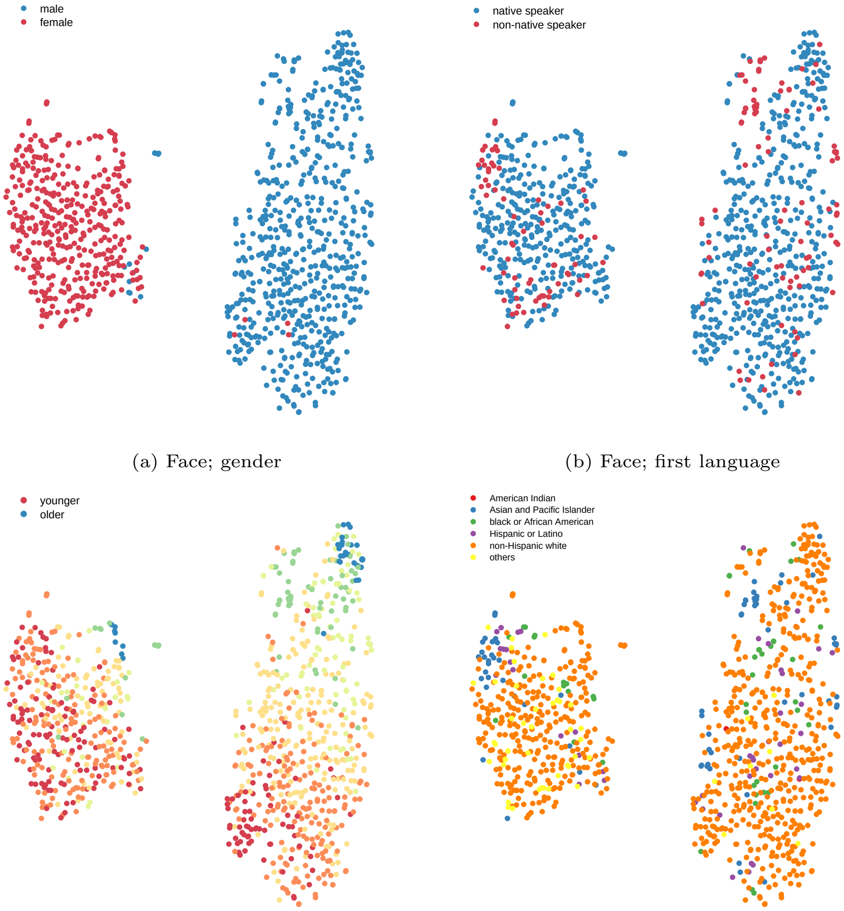
(a) Voice; gender

(b) Voice; first language

(c) Voice; age

(d) Voice; ethnicity

Fig. A.6. The t-SNE visualization of the *voice* representations of VoxCeleb test samples. 1,000 random samples are drawn from the test set and shown with four demographic attributes. (c) The color code depicts continuous values, while the legend shows only the minimum and the maximum values; the rest encode categorical values. See the main paper for the t-SNE with the voice identity marked.



(a) Face; gender

(b) Face; first language

(c) Face; age

(d) Face; ethnicity

Fig. A.7. The t-SNE visualization of the *face* representations of VoxCeleb test samples. 1,000 random samples are drawn from the test set and shown with four demographic attributes. (c) The color code depicts continuous values, while the legend shows only the minimum and the maximum values; the rest encode categorical values. See the main paper for the t-SNE with the face identity marked.

A.2.2 Further Discussions

Comparisons to binary classification. We experimented with a classification network inspired by the "L^3 network" [1], an audiovisual correlation classifier, and Nagrani et al.'s model [3], and trained to do binary classification: given a face and a voice, whether or not the two belong to the same identity. Here we detail the construction of the network. The classification network shares the same subnetworks as our architecture based on the triplet loss, but the two 512-d feature vectors average-pooled from the `conf5_3` and `conv6` of VGG16 and SoundNet, respectively, are concatenated to form a 1024-d vector, which is then fed to two 128-d fully-connected layers, in succession, followed by a 2-d fully-connected layer and the softmax activation. The class probability of the positive association is used as a score to measure the similarity of the face and the voice, hence for gauging the distances between a given voice (face) and two candidate faces (voices). The candidate with the higher similarity score is taken as the matching pair.

Dimensions of fully connected layers. We measured the test accuracy with varying dimensions of the fully connected layers (and thus the representation vectors), which is tabulated in Table A.2. While this did not have a significant influence on test accuracy, generally, narrower fully connected layers resulted in slightly better performance.

Further details on training. The batch size and the learning rate were chosen by grid search within the machine limit. Decaying learning rates and mining hard negative samples helped stabilize training and prevent from overfitting to training data, but did not contribute much to improve accuracy. The timing and

Table A.1. The analysis of encoded information in face and voice representations. This experiment is similar to Table 4 in the main paper, but with the triplet network trained with a face anchor subnetwork and positive and negative voice subnetworks. (In the main paper, we report the performance with the triplet network trained with a voice anchor subnetwork and positive and negative face subnetworks.) As a probe task, we use the attribute classification task. We report the mean average precision (mAP) with 99% confidence intervals (CI) obtained from 20 trials of holdout cross validations. We mark the values having confidence intervals that overlap with a random chance with a 5% margin, i.e., $50 \pm 5\%$, in red. In cases where the performance is less than or equal to random chance, it is suggested that the representation is not distinctive enough for the classification task. Note that during representation learning, no attribute information was seen by the network.

Modality		Gender	Fluency	Age					Ethnicity					
				<30	30s	40s	50s	≥60	1	2	3	4	5	6
Face repr.	mAP	98.7	67.9	71.9	68.4	57.6	63.1	81.4	90.3	79.6	81.9	71.1	67.8	74.3
	CI	±2.6	±4.1	±9.9	±3.7	±3.8	±7	±3.8	±4.3	±6	±5.5	±5.5	±6.9	±5.5
Voice repr.	mAP	93.1	58.2	65.7	56.7	52.7	56.7	62.9	94.5	71.1	58.4	61.1	55.8	68.5
	CI	±1.8	±3.4	±3.9	±3.6	±1.4	±3.7	±5.9	±2	±7.1	±4.6	±2.7	±4.6	±9.3

Table A.2. Test accuracy with varying fully connected layer dimensions (and thus our representation dimensions). For smaller dimensions, the last convolutional layer of each subnetwork is average-pooled globally before fed to the first fully-connected layer; for the dimensions larger than the filter dimension of the last convolutional layer (512-d), it was average-pooled with the factor of 2 along each non-singleton spatial dimension.

Experiments	Global spatial pooling		2× spatial pooling		
	128-d	512-d	1024-d	2048-d	4096-d
–	78.2%	77.4%	77.9%	77.7%	77.6%
G/E/F/A	59.0%	57.6%	58.5%	58.1%	58.2%

amount of decaying were set empirically. A standard data augmentation scheme was used optionally: face images are randomly cropped around the face region by −40% to 20%, rotated for a random angle between ±15°, and horizontally flipped randomly. Negative cropping means including more background. Image brightness and contrast as well as audio volume are jittered up to ±20%. We trained our network both with and without data augmentation under the same setup outlined below, but did not find significant difference in performance. This could be due to the large sample size (over 100k) and great diversity of the VoxCeleb dataset we used for training. Our model was implemented using TensorFlow and trained on an NVIDIA Titan X (Pascal) with 12 Gb RAM. Training typically takes less than a day on a single GPU.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV. pp. 609–617 (2017)
2. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. JMLR **9**, 2579–2605 (2008)
3. Nagrani, A., Albanie, S., Zisserman, A.: Seeing voices and hearing faces: Cross-modal biometric matching. In: CVPR. pp. 8427–8436 (2018)
4. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: ICCV. pp. 1034–1041 (2009)