A. Further Ablation Experiments

What is the minimum duration of audio segment for reliable performance?

We ran experiments using the best static model with audio segments of $2/3, 1, 2, 3$ and 4 seconds extracted at both train and test time[4] (results are reported in Table 4). We were unable to achieve convergence with $2/3$s, suggesting the minimum duration to be between $2/3$s and 1s. For longer durations, we found our method to be relatively robust to variations in length.

Duration (s)	2/3	1	2	3	4
Accuracy (%)	52.8 (chance= 50)	77.4	78.6	79.2	79.1

Table 4: Results on the static matching task using different audio segment lengths.

Which factor (gender, age, nationality) has the highest effect on performance?

We conducted an experiment in which test triplets are matched on each factor separately, e.g. for age (A), all elements of a triplet have a similar age. We report the following numbers (% acc.) for the best static model: original performance 81.0, nationality (N) 78.2, age (A) 75.0, gender (G) 65.2, matching all three (GNA) 63.9. Matching on gender has the greatest effect on performance. This could be because (i) the dataset [33] does not have enough age and nationality variation (as demonstrated in appendix B); and (ii) gender is more discriminative a factor for the task (especially given that nationality is not always indicative of accent). This is also supported by the visual results in Figure 9, which show that the most highly ranked pairs classified correctly are those of different genders.

B. Test set statistics

As described in section 4 of the paper, the nationalities of the speakers in the test set were obtained by crawling Wikipedia. Figure 6 illustrates the distribution of nationality and gender across this data. In order to create the more challenging *GNA-var removed* dataset, we use only US nationals (of both genders) between the ages of 30-50 years old. The age estimates were obtained manually. The final *GNA-var removed* dataset has 110 speakers.

C. Amazon Mechanical Turk Study

In this section, we describe in more detail the experimental methodology employed to establish a benchmark for human performance on the test set, described in section 5.2. A pool of Amazon Mechanical Turk (AMT) workers were presented with a set of matching problems, each consisting of two face images and a single audio segment. In each matching problem, they were asked to choose which of the faces corresponded to the person speaking. To ensure that workers were actually listening to the audio segments, a result could only be submitted once the audio sample had been played (achieved by deactivating the selection buttons). Workers could listen to the audio samples as many times as required.

In total 500 test triplets were selected, and shown to twenty workers in batches of five (in order to prevent worker fatigue). To avoid the workers 'learning' the face-voice pairings, batches were chosen to ensure the same speaker was not present in the same batch. The accuracy was then computed for each worker over all the triplets that they labelled, and averaged across all workers to produce an estimate of human accuracy. If a worker achieved an accuracy below 40%, their results were discarded. To obtain a measure of variance, the mean standard deviation of worker accuracy on each test triplet was calculated and was found to be 2.55%. A screenshot of the webpage seen by workers is shown in Figure 7.

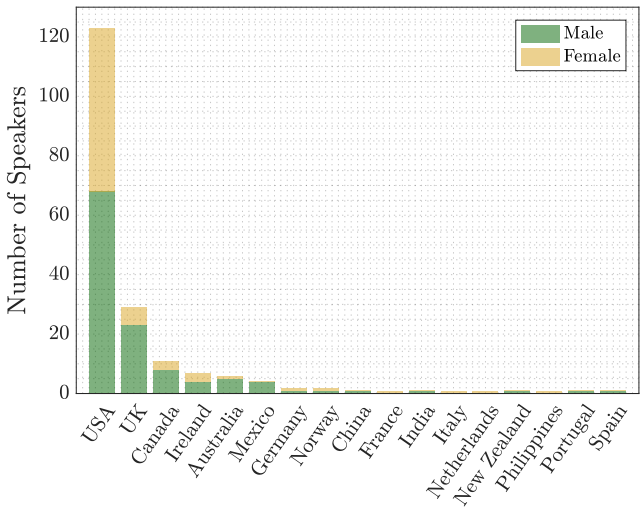

Figure 6: Distribution of nationalities of the speakers in the test set.

D. Network Architecture Details

The filter and output sizes for both voice and face subnetworks can be seen in figure 8.

[4]To investigate lengths up to 4s and perform a fair comparison, we restricted the dataset to speech segments that were 5s or longer (74% of the total dataset). This ensures the size of the dataset is fixed for each experiment. At test time, sub-segments are densely sampled with the given duration and predictions are averaged.

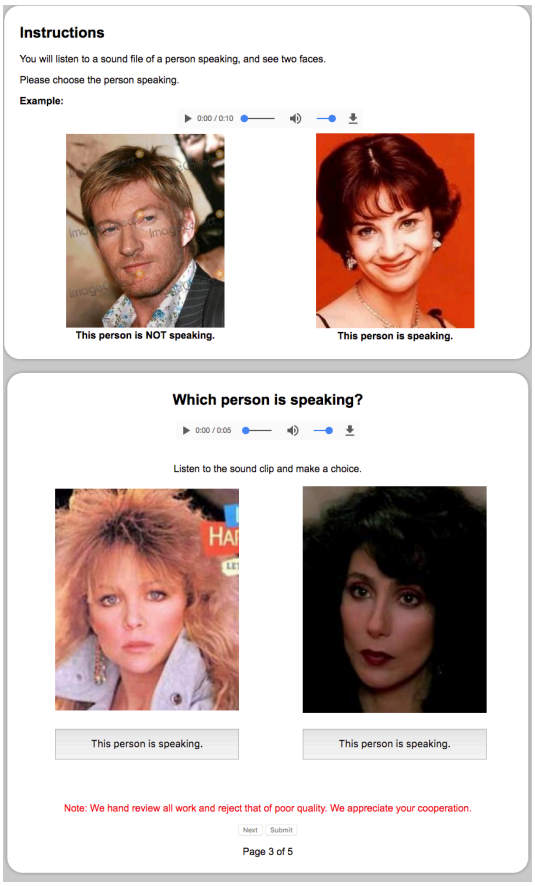

Figure 7: **Screenshot of the webpage shown to workers**. (Top) Upon starting each batch, each worker was provided with instructions and an illustrative example. (Bottom) The interface for submitting worker predictions.

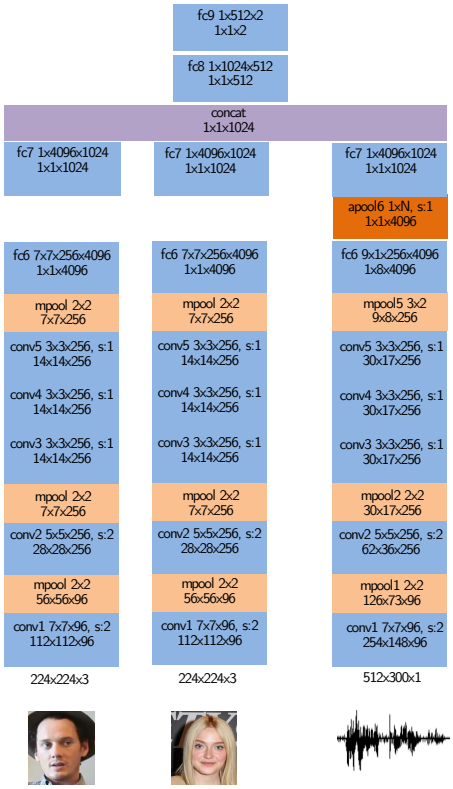

Figure 8: Static architecture for forced matching between two faces and one voice segment (V-F formulation). Note how average pooling is used in the voice subnetwork to deal with variable length speech segments. The value of N in `apool6` changes according to the size of the speech segment input. (Output sizes up till `apool6` are shown for an input speech segment of three seconds, for which $N = 8$ in `apool6` .)

E. Salient regions

Given the strong performance of the static image model on the challenging *GNA-var removed* evaluation set (as discussed in section 6 of the paper), we would like to gain some insight into how the network is accomplishing the task. The interpretation of the model class of deep neural networks remains a challenging topic and an area of active research (see e.g. [30, 32, 54]). One approach to understanding the decision making process of the network is through region saliency, in which the goal is to identify regions in input space which have exerted maximal influence on a classification decision. Here, we employ the *Excitation Backprop* method introduced in [55] to find discriminative regions in the face inputs[5]. Specifically, we use the *contrastive attention* technique introduced in [55] to visualise saliency following the `relu3` layer in the face streams (we found that higher layers were less informative, typically producing a response covering the extent of the face). The resulting visualisations are shown for samples from the *GNA-var removed* test set in Figure 10. We observe that the model often finds highly localised regions in the lower half of the face particularly salient for voice matching (first two rows). However, we also found that in certain cases, it is strongly influenced by a region of greater spatial extent, including the nose and cheeks (third row), or combinations of distinctive features, such as the eyes and mouth (fourth row). Rather than depending on a single consistent feature to solve the task, it therefore seems that the network has learned to draw selectively from a range of signals to classify voices robustly.

[5]Unfortunately voice data, which is consumed by the model in the form of a spectrogram, is less amenable to visual interpretation.

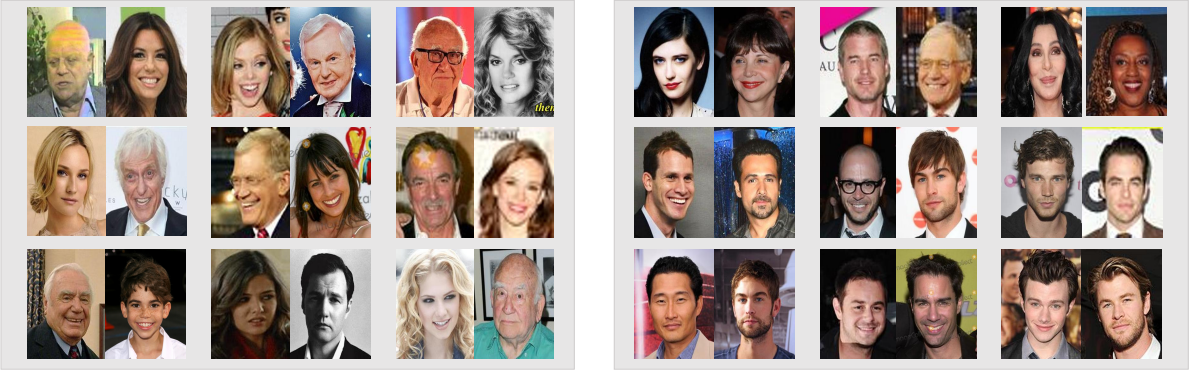

Figure 9: Examples of the top ranked face pairs that were classified correctly using a single voice segment (left panel) and the bottom ranked classified incorrectly (right panel) on the static test set. From the images on the left, it is clear that the model finds it easier to distinguish between faces of different gender and age.

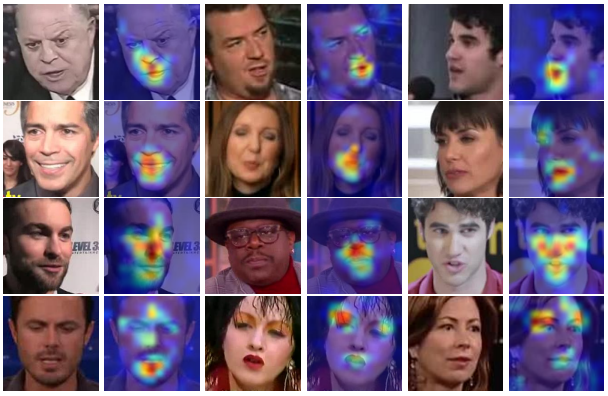

Figure 10: **Salient facial regions for voice classification** - Each example depicts a sampled input face (left) and its corresponding saliency map for the voice matching task (right). The first two rows show highly localised discriminative regions in the lower portion of the face. The final two rows show a more distributed response and usage of other features, particularly eyes. See text for further discussion.