Speech2Face: Learning the Face Behind a Voice

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

| Citation       | Oh, Tae-Hyun et al. “Speech2Face: Learning the Face Behind a Voice.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2019, Long Beach, California, Institute of Electrical and Electronics Engineers, January 2020. © 2019 IEEE |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| As Published   | http://dx.doi.org/10.1109/cvpr.2019.00772                                                                                                                                                       |
| Publisher      | Institute of Electrical and Electronics Engineers (IEEE)                                                                                                                                          |
| Version        | Original manuscript                                                                                                                                                                            |
| Citable link   | https://hdl.handle.net/1721.1/130277                                                                                                                                                           |
| Terms of Use   | Creative Commons Attribution-Noncommercial-Share Alike                                                                                                                                          |
| Detailed Terms | http://creativecommons.org/licenses/by-nc-sa/4.0/                                                                                                                                              |
Speech2Face: Learning the Face Behind a Voice

Tae-Hyun Oh∗† | Tali Dekel* | Changil Kim∗† | Inbar Mosseri
William T. Freeman† | Michael Rubinstein | Wojciech Matusik†

†MIT CSAIL

Abstract

How much can we infer about a person's looks from the way they speak? In this paper, we study the task of reconstructing a facial image of a person from a short audio recording of that person speaking. We design and train a deep neural network to perform this task using millions of natural Internet/YouTube videos of people speaking. During training, our model learns voice-face correlations that allow it to produce images that capture various physical attributes of the speakers such as age, gender and ethnicity. This is done in a self-supervised manner, by utilizing the natural co-occurrence of faces and speech in Internet videos, without the need to model attributes explicitly. We evaluate and numerically quantify how—and in what manner—our Speech2Face reconstructions, obtained directly from audio, resemble the true face images of the speakers.

1. Introduction

When we listen to a person speaking without seeing his/her face, on the phone, or on the radio, we often build a mental model for the way the person looks [25, 45]. There is a strong connection between speech and appearance, part of which is a direct result of the mechanics of speech production: age, gender (which affects the pitch of our voice), the shape of the mouth, facial bone structure, thin or full lips—all can affect the sound we generate. In addition, other voice-appearance correlations stem from the way in which we talk: language, accent, speed, pronunciations—such properties of speech are often shared among nationalities and cultures, which can in turn translate to common physical features [12].

Our goal in this work is to study to what extent we can infer how a person looks from the way they talk. Specifically, from a short input audio segment of a person speaking, our method directly reconstructs an image of the person’s face in a canonical form (i.e., frontal-facing, neutral expression). Fig. 1 shows sample results of our method. Obviously, there is no one-to-one matching between faces and voices. Thus, our goal is not to predict a recognizable image of the exact face, but rather to capture dominant facial traits of the person that are correlated with the input speech.

We design a neural network model that takes the complex spectrogram of a short speech segment as input and predicts a feature vector representing the face. More specifically, face information is represented by a 4096-D feature that is extracted from the penultimate layer (i.e., one layer prior to the classification layer) of a pre-trained face recognition network [40]. We decode the predicted face feature into a canonical image of the person’s face using a separately-trained reconstruction model [10]. To train our model, we use the AVSpeech dataset [14], comprised of millions of video segments from YouTube with more than 100,000 different people speaking. Our method is trained in a self-supervised manner, i.e., it simply uses the natural co-occurrence of speech and faces in videos, not requiring additional information, e.g., human annotations.

∗The three authors contributed equally to this work.
Correspondence: taehyun@csail.mit.edu
Supplementary material (SM): https://speech2face.github.io
We are certainly not the first to attempt to infer information about people from their voices. For example, predicting age and gender from speech has been widely explored [53, 18, 16, 7, 50]. Indeed, one can consider an alternative approach to attaching a face image to an input voice by first predicting some attributes from the person’s voice (e.g., their age, gender, etc., [53]), and then either fetching an image from a database that best fits the predicted set of attributes, or using the attributes to generate an image [52]. However, this approach has several limitations. First, predicting attributes from an input signal relies on the existence of robust and accurate classifiers and often requires ground truth labels for supervision. For example, predicting age, gender or ethnicity from speech requires building classifiers specifically trained to capture those properties. More importantly, this approach limits the predicted face to resemble only a predefined set of attributes.

We aim at studying a more general, open question: what kind of facial information can be extracted from speech? Our approach of predicting the target visual appearance (e.g., a face image) directly from speech allows us to explore this question without being restricted to predefined facial traits. Specifically, we show that our reconstructed face images can be used as a proxy to convey the visual properties of the person including age, gender and ethnicity. Beyond these dominant features, our reconstructions reveal non-negligible correlations between craniofacial features [31] (e.g., nose structure) and voice. This is achieved with no prior information or the existence of accurate classifiers for these types of fine geometric features. In addition, we believe that predicting face images directly from voice may support useful applications, such as attaching a representative face to phone/video calls based on the speaker’s voice.

To our knowledge, our work is the first to explore a generic (speaker independent) model for reconstructing face images directly from speech. We test our model on various speakers and numerically evaluate different aspects of our reconstructions including: how well a true face image can be retrieved based solely on an audio query; and how well our reconstructed face images agree with the true face images (unknown to the method) in terms of age, gender, ethnicity, and various craniofacial measures and ratios.

2. Ethical Considerations

Although this is a purely academic investigation, we feel that it is important to explicitly discuss in the paper a set of ethical considerations due to the potential sensitivity of facial information.

Privacy. As mentioned, our method cannot recover the true identity of a person from their voice (i.e., an exact image of their face). This is because our model is trained to capture visual features (related to age, gender, etc.) that are common to many individuals, and only in cases where there is strong evidence to connect those visual features with vocal/speech attributes in the data (see “voice-face correlations” below). As such, the model will only produce average-looking faces, with characteristic visual features that are correlated with the input speech. It will not produce images of specific individuals.

Voice-face correlations and dataset bias. Our model is designed to reveal statistical correlations that exist between facial features and voices of speakers in the training data. The training data we use is a collection of educational videos from YouTube [14], and does not represent equally the entire world population. Therefore, the model—as is the case with any machine learning model—is affected by this uneven distribution of data.

More specifically, if a set of speakers might have vocal-visual traits that are relatively uncommon in the data, then the quality of our reconstructions for such cases may degrade. For example, if a certain language does not appear in the training data, our reconstructions will not capture well the speech, for example hair color or style. However, if many speakers in the training set who speak in a similar way (e.g., in the same language)
We use a pre-trained face recognition network to transfer visual reconstruction from audio. Various methods of audio-visual cross-modal learning have been used to disambiguate voiced and unvoiced faces have been shown to contain enough information to reconstruct the visual modality to the sound modality, using unlabeled videos. Similarly, Castrejon et al. [8] designed a shared audio-visual representation that is agnostic of the modality. Such learned audio-visual representations have been used for cross-modal retrieval [38, 39, 46], sound source localization [42, 5, 37], and sound source separation [54, 14].

4. Speech2Face (S2F) Model

The large variability in facial expressions, head poses, occlusions, and lighting conditions in natural face images makes the design and training of a Speech2Face model non-trivial. For example, a straightforward approach of regressing from input speech to image pixels does not work; such a model has to learn to factor out many irrelevant variations in the data and to implicitly extract a meaningful internal representation of faces—a challenging task by itself.

To sidestep these challenges, we train our model to regress to a low-dimensional intermediate representation of the face. More specifically, we utilize the VGG-Face model, a face recognition model pre-trained on a large-scale face dataset [40], and extract a 4096-D face feature from the penultimate layer of the network. These face features were shown to contain enough information to reconstruct the corresponding face images while being robust to many of the aforementioned variations [10].

Our Speech2Face pipeline, illustrated in Fig. 2, consists of two main components: 1) a voice encoder, which takes a complex spectrogram of speech as input, and predicts a low-dimensional face feature that would correspond to the associated face; and 2) a face decoder, which takes as input the face feature and produces an image of the face in a canonical form (frontal-facing and with neutral expression). During training, the face decoder is fixed, and we train only the voice encoder that predicts the face feature. The voice encoder is a model we designed and trained, while we used a face decoder model proposed by Cole et al. [10]. We now describe both models in detail.

Voice encoder network. Our voice encoder module is a convolutional neural network that turns the spectrogram of a short input speech into a pseudo face feature, which is subsequently fed into the face decoder to reconstruct the face image (Fig. 2). The architecture of the voice encoder is summarized in Table 1. The blocks of a convolution layer, ReLU,
Figure 3. **Qualitative results on the AVSpeech test set.** For every example (triplet of images) we show: (left) the original image, i.e., a representative frame from the video cropped around the speaker's face; (middle) the frontalized, lighting-normalized face decoder reconstruction from the VGG-Face feature extracted from the original image; (right) our Speech2Face reconstruction, computed by decoding the predicted VGG-Face feature from the audio. In this figure, we highlight successful results of our method. Some failure cases are shown in Fig. 12, and more results (including the input audio for all the examples) can be found in the SM.
Table 1. Voice encoder architecture. The input spectrogram dimensions are 598 × 257 (time × frequency) for a 6-second audio segment (which can be arbitrarily long), with two input channels in the table corresponding to the spectrogram’s real and imaginary components.

| Layer | Input | CONV | RELU | CONV | RELU | MAXPOOL | CONV | RELU | MAXPOOL | CONV | RELU | MAXPOOL | CONV | RELU | MAXPOOL | CONV | AVGPOOL | RELU | FC | FC |
|-------|-------|------|------|------|------|---------|------|------|---------|------|------|---------|------|------|---------|------|---------|------|----|----|
| Channels | 2 | 64 | 64 | 128 | – | 128 | – | 128 | – | 256 | – | 512 | 512 | 512 | – | 4096 | 4096 |
| Stride | – | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 |
| Kernel size | – | 3 × 4 | 3 × 4 | 4 × 4 | 2 × 1 | 4 × 4 | 2 × 1 | 4 × 4 | 2 × 1 | 4 × 4 | 2 × 1 | 4 × 4 | 4 × 4 | 4 × 4 | 4 × 4 | 1 × 1 | 1 × 1 | 1 × 1 |

Figure 4. Facial attribute evaluation. (a) confusion matrices (with row-wise normalization) comparing the classification results on our Speech2Face image reconstructions (S2F) and those obtained from the original images for gender, age, and ethnicity; the stronger diagonal tendency the better performance. Ethnicity performance in (a) appears to be biased due to uneven distribution of the training set shown in (b).

and batch normalization [23] alternate with max-pooling layers, which pool along only the temporal dimension of the spectrograms, while leaving the frequency information carried over. This is intended to preserve more of the vocal characteristics, since they are better contained in the frequency content, whereas linguistic information usually spans longer time duration [22]. At the end of these blocks, we apply average pooling along the temporal dimension. This allows us to efficiently aggregate information over time and makes the model applicable to input speech of varying duration. The pooled features are then fed into two fully-connected layers to produce a 4096-D face feature.

Face decoder network. The goal of the face decoder is to reconstruct the image of a face from a low-dimensional face feature. We opt to factor out any irrelevant variations (pose, lighting, etc.), while preserving the facial attributes. To do so, we use the face decoder model of Cole et al. [10] to reconstruct a canonical face image. We train this model using the same face features extracted from the VGG-Face model as input to the face decoder. This model is trained separately and kept fixed during the voice encoder training.

Training. Our voice encoder is trained in a self-supervised manner, using the natural co-occurrence of a speaker’s speech and facial images in videos. To this end, we use the AVSpeech dataset [14], a large-scale “in-the-wild” audio-visual dataset of people speaking. A single frame containing the speaker’s face is extracted from each video clip and fed to the VGG-Face model to extract the 4096-D feature vector, \( v_f \). This serves as the supervision signal for our voice encoder—the feature, \( v_s \), of our voice encoder is trained to predict \( v_f \).

A natural choice for the loss function would be the L1 distance between the features: \( \|v_f - v_s\|_1 \). However, we found that the training undergoes slow and unstable progression with this loss alone. To stabilize the training, we introduce additional loss terms, motivated by Castrojen et al. [8]. Specifically, we additionally penalize the difference in the activation of the last layer of the face encoder, \( f_{VGG}: \mathbb{R}^{4096} \rightarrow \mathbb{R}^{2622} \), i.e., fc8 of VGG-Face, and that of the first layer of the face decoder, \( f_{dec}: \mathbb{R}^{4096} \rightarrow \mathbb{R}^{1000} \), which are pre-trained and fixed during training the voice encoder. We feed both our predictions and the ground truth face features to these layers to calculate the losses. The final loss is:

\[
L_{\text{total}} = \| f_{\text{dec}}(v_f) - f_{\text{dec}}(v_s) \|_1 + \lambda_1 \left\| \frac{v_f}{\|v_f\|} - \frac{v_s}{\|v_s\|} \right\|^2_2 + \lambda_2 L_{\text{distill}}(f_{VGG}(v_f), f_{VGG}(v_s)),
\]

where \( \lambda_1 = 0.025 \) and \( \lambda_2 = 200 \). \( \lambda_1 \) and \( \lambda_2 \) are tuned such that the gradient magnitude of each term with respect to \( v_s \) are within a similar scale at an early iteration (we measured at the 1000th iteration). The knowledge distillation loss \( L_{\text{distill}}(a, b) = -\sum_i p_i(a) \log p_i(b) \), where \( p_i(a) = \frac{\exp(\alpha_i/T)}{\sum_j \exp(\alpha_j/T)} \), is used as an alternative of the cross-entropy loss, which encourages the output of a network to approximate the output of another [19]. \( T = 2 \) is used as
recommended by the authors, which makes the activation smoother. We found that enforcing similarity over these additional layers stabilized and sped up the training process, in addition to a slight improvement in the resulting quality.

**Implementation details.** We use up to 6 seconds of audio taken from the beginning of each video clip in AVSpeech. If the video clip is shorter than 6 seconds, we repeat the audio such that it becomes at least 6-seconds long. The audio waveform is resampled at 16 kHz and only a single channel is used. Spectrograms are computed similarly to Ephrat et al. [14] by taking STFT with a Hann window of 25 mm, the hop length of 10 ms, and 512 FFT frequency bands. Each complex spectrogram $S$ subsequently goes through the power-law compression, resulting $\text{sgn}(S)|S|^{0.3}$ for real and imaginary independently, where $\text{sgn}(\cdot)$ denotes the signum. We run the CNN-based face detector from Dlib [28], crop the face regions from the frames, and resize them to $224 \times 224$ pixels. The VGG-Face features are computed from the resized face images. The computed spectrogram and VGG-Face feature of each segment are collected and used for training. The resulting training and test sets include 1.7 and 0.15 million spectra–face feature pairs, respectively. Our network is implemented in TensorFlow and optimized by ADAM [29] with $\beta_1 = 0.5, \epsilon = 10^{-4}$, the learning rate of 0.001 with the exponentially decay rate of 0.95 at every 10,000 iterations, and the batch size of 8 for 3 epochs.

5. Results

We test our model both qualitatively and quantitatively on the AVSpeech dataset [14] and the VoxCeleb dataset [35]. Our goal is to gain insights and to quantify how—and in which manner—our Speech2Face reconstructions resemble the true face images.

Qualitative results on the AVSpeech test set are shown in Fig. 3. For each example, we show the true image of the speaker for reference (unknown to our model), the face reconstructed from the face feature (computed from the true image) by the face decoder (Sec. 4), and the face reconstructed from a 6-seconds audio segment of the person’s speech, which is our Speech2Face result. While looking somewhat like average faces, our Speech2Face reconstructions capture rich physical information about the speaker, such as their age, gender, and ethnicity. The predicted images also capture additional properties like the shape of the face or head (e.g., elongated vs. round), which we often find consistent with the true appearance of the speaker; see the last two rows in Fig. 3 for instance.

5.1. Facial Features Evaluation

We quantify how well different facial attributes are being captured in our Speech2Face reconstructions and test different aspects of our model.

**Demographic attributes.** We use Face++ [15], a leading commercial service for computing facial attributes. Specifically, we evaluate and compare age, gender, and ethnicity, by running the Face++ classifiers on the original images and our Speech2Face reconstructions. The Face++ classifiers return either “male” or “female” for gender, a continuous number for age, and one of the four values, “Asian”, “black”, “India”, or “white”, for ethnicity. What we refer to as “white” in our work corresponds to “Asian” in the Face++ terminology.

Fig. 4(a) shows confusion matrices for each of the attributes, comparing the attributes inferred from the original images with those inferred from our Speech2Face reconstructions (S2F). See the supplementary material for similar evaluations of our face-decoder reconstructions from the images (F2F). As can be seen, for age and gender the classification results are highly correlated. For gender, there is an agreement of 94% in male/female labels between the true images and our reconstructions from speech. For ethnicity, there is a good correlation on the “white” and “Asian”, but we observe less agreement on “India” and “black”. We

---

1. We directly refer to the Face++ labels, which are not our terminology.
believe this is because those classes have a smaller representation in the data (see statistics we computed on AVSpeech in Fig. 4(b)). The performance can potentially be improved by leveraging the statistics to balance the training data for the voice encoder model, which we leave for future work.

**Craniofacial attributes.** We evaluated craniofacial measurements commonly used in the literature, for capturing ratios and distances in the face [31]. For each such measurement, we computed the correlation between F2F (Fig. 5(a)), and our corresponding S2F reconstructions (Fig. 5(b)). Face landmarks were computed using the DEST library [1]. Note that this evaluation is made possible because we are working with normalized faces (neutral expression, frontal-facing), thus differences between the facial landmarks’ positions reflect geometric craniofacial changes. Fig. 5(c) shows the Pearson correlation coefficient for several measures, computed over 1,000 random samples from the AVSpeech test set. As can be seen, there is statistically significant (i.e., $p < 0.001$) positive correlation for several measurements. In particular, the highest correlation is measured for the nasal index (0.38) and nose width (0.35), the features indicative of nose structures that may affect a speaker’s voice.

**Feature similarity.** We test how well a person can be recognized from the face features predicted from speech. We first directly measure the cosine distance between our predicted features and the true ones obtained from the original face image of the speaker. Table 2 shows the average error over 5,000 test images, for the predictions using 3s and 6s audio segments. The use of longer audio clips exhibits consistent improvement in all error metrics; this further evidences the qualitative improvement we observe in Fig. 6.

![Figure 6. The effect of input audio duration.](image)

We further evaluated how accurately we can retrieve the true speaker from a database of face images. To do so, we take the speech of a person to predict the feature using our Speech2Face model, and query it by computing its distances to the face features of all face images in the database. We report the retrieval performance by measuring the recall at $K$, i.e., the percentage of time the true face is retrieved within the top-$K$ results. We used a database of 5,000 images for this experiment; see Fig. 7 for qualitative results. The higher the better. Random chance is presented as a baseline.

![Figure 7. S2F→Face retrieval examples.](image)

**Table 2. Feature similarity.** We measure the similarity between our features predicted from speech and the corresponding face features computed on the true images of the speakers. We report average cosine, $L_2$ and $L_1$ distances over 5000 random samples from the AVSpeech test set, using 3- and 6-second audio segments.

| Length | cos (deg) | $L_2$ | $L_1$ |
|--------|-----------|-------|-------|
| 3 seconds | 48.43 ± 6.01 | 0.19 ± 0.03 | 9.81 ± 1.74 |
| 6 seconds | 45.75 ± 5.09 | 0.18 ± 0.02 | 9.42 ± 1.54 |

**Table 3. S2F→Face retrieval performance.** We measure retrieval performance by recall at $K$ ($R@K$, in %), which indicates the chance of retrieving the true image of a speaker within the top-$K$ results. We used a database of 5,000 images for this experiment; see Fig. 7 for qualitative results. The higher the better. Random chance is presented as a baseline.
Figure 8. **Comparison to a pixel loss.** The results obtained with an $L_1$ loss on the output image and our full loss (Eq. 1) are shown after 300k and 500k training iterations (indicating convergence).

Figure 9. **Training convergence patterns.** BN denotes batch normalization. The red and green curves are obtained by using 3- and 6-second audio clips as input during training, respectively (dashed line: training loss; solid line: validation loss). The face thumbnails show reconstructions from models trained with and without BN.

which demonstrate the consistent facial characteristics that are being captured by our predicted face features.

**t-SNE visualization for learned feature analysis.** To gain more insights on our predicted features, we present 2-D t-SNE plots [49] of the features in the SM.

### 5.2. Ablation Studies

**Comparisons with a direct pixel loss.** Fig. 8 shows qualitative comparisons between the model trained with our full loss (Eq. 1) and the same model trained with only an image loss, i.e., an $L_1$ loss between pixel values on the decoded image layer (with the decoder fixed). The model trained with the image loss results in lower facial image quality and fewer facial variations. Our loss, measured at an early layer of the face decoder, allows for better supervision and leads to faster training and higher quality results.

**The effect of audio duration and batch normalization.**

We tested the effect of the duration of the input audio during both the train and test stages. Specifically, we trained two models with 3- and 6-second speech segments. We found that during the training time, the audio duration has an only subtle effect on the convergence speed, without much effect on the overall loss and the quality of reconstructions (Fig. 9). However, we found that feeding longer speech as input at test time leads to improvement in reconstruction quality, that is, reconstructed faces capture the personal attributes better, regardless of which of the two models are used. Fig. 6 shows several qualitative comparisons, which are also consistent with the quantitative evaluations in Tables 2 and 3.

Fig. 9 also shows the training curves w/ and w/o Batch Normalization (BN). As can be seen, without BN the reconstructed faces converge to an average face. With BN the results contain much richer facial information.

**Additional observations and limitations.** In Fig. 10, we infer faces from different speech segments of the same person, taken from different parts within the same video, and from a different video, in order to test the stability of our Speech2Face reconstruction. The reconstructed face images are consistent within and between the videos. We show more such results in the SM.

To qualitatively test the effect of language and accent, we probe the model with an Asian male example speaking the same sentence in English and Chinese (Fig. 11(a)). While having the same reconstructed face in both cases would be
ideal, the model inferred different faces based on the spoken language. However, in other examples, e.g., Fig. 11(b), the model was able to successfully factor out the language, reconstructing a face with Asian features even though the girl was speaking in English with no apparent accent (the audio is available in the SM). In general, we observed mixed behaviors and a more thorough examination is needed to determine to which extent the model relies on language.

More generally, the ability to capture the latent attributes from speech, such as age, gender, and ethnicity, depends on several factors such as accent, spoken language, or voice pitch. Clearly, in some cases, these vocal attributes would not match the person’s appearance. Several such typical speech-face mismatch examples are shown in Fig. 12.

5.3. Speech2cartoon

Our face images reconstructed from speech may be used for generating personalized cartoons of speakers from their voices, as shown in Fig. 13. We use Gboard, the keyboard app available on Android phones, which is also capable of analyzing a selfie image to produce a cartoon-like version of the face [32]. As can be seen, our reconstructions capture the facial attributes well enough for the app to work.

![Figure 12. Example failure cases. (a) High-pitch male voice, e.g., of kids, may lead to a face image with female features. (b) Spoken language does not match ethnicity. (c-d) Age mismatches.](image)

![Figure 13. Speech-to-cartoon. Our reconstructed faces from audio (b) can be re-rendered as cartoons (c) using existing tools, such as the personalized emoji app available in Gboard, the keyboard app in Android phones [32]. (a) The true images of the person are shown for reference.](image)

Such cartoon re-rendering of the face may be useful as a visual representation of a person during a phone or a video-conferencing call, when the person’s identity is unknown or the person prefers not to share his/her picture. Our reconstructed faces may also be used directly, to assign faces to machine-generated voices used in home devices and virtual assistants.

6. Conclusion

We have presented a novel study of face reconstruction directly from the audio recording of a person speaking. We address this problem by learning to align the feature space of speech with that of a pre-trained face decoder using millions of natural videos of people speaking. We have demonstrated that our method can predict plausible faces with the facial attributes consistent with those of real images. By reconstructing faces directly from this cross-modal feature space, we validate visually the existence of cross-modal biometric information postulated in previous studies [27, 34]. We believe that generating faces, as opposed to predicting specific attributes, may provide a more comprehensive view of voice-face correlations and can open up new research opportunities and applications.

Acknowledgment The authors would like to thank Suwon Shon, James Glass, Forrester Cole and Dilip Krishnan for helpful discussion. T.-H. Oh and C. Kim were supported by QCRI-CSAIL Computer Science Research Program at MIT.

References

[1] One Millisecond Deformable Shape Tracking Library (DEST). https://github.com/cheind/dest.
[2] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In ACM Multimedia Conference (MM), 2018.
[3] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In International Conference on Machine Learning (ICML), 2013.
[4] R. Arandjelovic and A. Zisserman. Look, listen and learn. In IEEE International Conference on Computer Vision (ICCV), 2017.
[5] R. Arandjelovic and A. Zisserman. Objects that sound. In European Conference on Computer Vision (ECCV), 2018.
[6] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In Advances in Neural Information Processing Systems (NIPS), 2016.
[7] M. H. Bahari and H. Van Hamme. Speaker age estimation and gender detection based on supervised non-negative matrix factorization. In IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications, 2011.
[8] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba. Learning aligned cross-modal representations from weakly aligned data. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[44] S. Shon, T.-H. Oh, and J. Glass. Noise-tolerant audio-visual online person verification using an attention-based neural network fusion. CoRR, abs/1811.10813, 2018. 3

[45] H. M. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey. Matching novel face and voice identity using static and dynamic facial images. Attention, Perception, & Psychophysics, 78(3):868–879, 2016. 1

[46] M. Solèr, J. C. Bazin, O. Wang, A. Krause, and A. Sorkine-Hornung. Suggesting sounds for images from video collections. In European Conference on Computer Vision Workshops, 2016. 3

[47] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. ACM Transactions on Graphics (SIGGRAPH), 36(4):95, 2017. 3

[48] S. L. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. K. Hodgins, and I. A. Matthews. A deep learning approach for generalized speech animation. ACM Transactions on Graphics (SIGGRAPH), 36(4):93:1–93:11, 2017. 3

[49] L. van der Maaten and G. Hinton. Visualizing data using T-SNE. Journal of Machine Learning Research (JMLR), 9(Nov):2579–2605, 2008. 8

[50] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh. Disjoint mapping network for cross-modal matching of voices and faces. In International Conference on Learning Representations (ICLR), 2019. 2, 3

[51] O. Wiles, A. S. Koepke, and A. Zisserman. X2Face: A network for controlling face generation using images, audio, and pose codes. In European Conference on Computer Vision (ECCV), Springer, 2018. 3

[52] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In European Conference on Computer Vision (ECCV), Springer, 2016. 2, 3

[53] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak. Age estimation in short speech utterances based on LSTM recurrent neural networks. IEEE Access, 6:22524–22530, 2018. 2

[54] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. H. McDermott, and A. Torralba. The sound of pixels. In European Conference on Computer Vision (ECCV), Springer, 2018. 3