S^3: Speech, Script and Scene driven Head and Eye Animation

YIFANG PAN, University of Toronto, Canada and Jali Research, Canada
RISHABH AGRAWAL, Jali Research, Canada
KARAN SINGH, University of Toronto, Canada and Jali Research, Canada

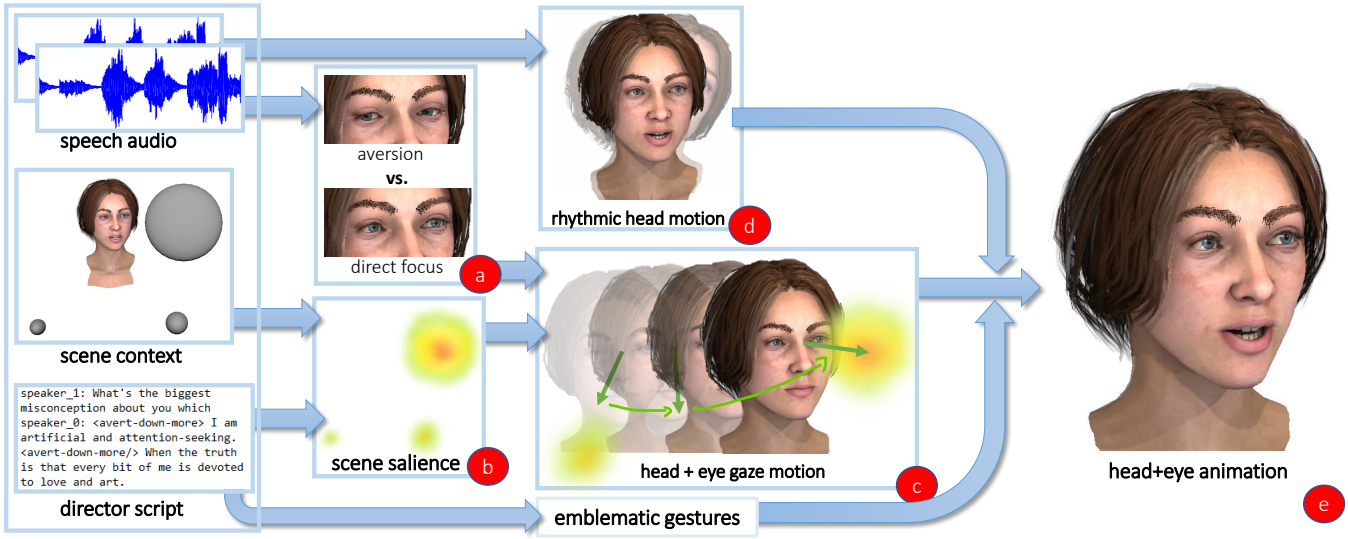


Fig. 1. Conversational audio, and a tagged transcript are aligned and diarized into separate streams. Speaker gaze during segments of speech are predicted as focused-on or averted-from a conversation partner (a). A 3D scene context defines a dynamic saliency map (b), which refines the predicted gaze transitions, into a set of 3D gaze trajectories (c). Speech audio generates rhythmic head motion (d), and it is used with other gestures to produce head+eye motion satisfying the gaze trajectories (e).

We present S^3, a novel approach to generating expressive, animator-centric 3D head and eye animation of characters in conversation. Given *speech* audio, a Directorial *script* and a cinematographic 3D *scene* as input, we automatically output the animated 3D rotation of each character's head and eyes. S^3 distills animation and psycho-linguistic insights into a novel modular framework for conversational gaze capturing: audio-driven rhythmic head motion; narrative script-driven emblematic head and eye gestures; and gaze trajectories computed from audio-driven gaze focus/aversion and 3D visual scene salience. Our evaluation is four-fold: we quantitatively validate our algorithm against ground truth data and baseline alternatives; we conduct a perceptual study showing our results to compare favourably to prior art; we present examples of animator control and critique of S^3 output; and present a large number of compelling and varied animations of conversational gaze.

CCS Concepts: • **Computing methodologies** → **Animation**.

Authors' addresses: Yifang Pan, University of Toronto, Canada and Jali Research, Canada, evan.pan@dgp.toronto.edu; Rishabh Agrawal, Jali Research, Canada, rishabhagrawal1995@outlook.com; Karan Singh, University of Toronto, Canada and Jali Research, Canada, karan@dgp.toronto.edu.

ACM 0730-0301/2024/7-ART47
https://doi.org/10.1145/3658172

Additional Key Words and Phrases: facial animation, head and eye control, gaze behavior, speech.

ACM Reference Format:
Yifang Pan, Rishabh Agrawal, and Karan Singh. 2024. **S^3**: Speech, Script and Scene driven Head and Eye Animation. *ACM Trans. Graph.* 43, 4, Article 47 (July 2024), 12 pages. https://doi.org/10.1145/3658172

1 INTRODUCTION

Our head, through rhythmic gestural motion, and eyes, our "windows to the soul", through subtle spatio-temporal changes in gaze, play a quintessential role in expressive, non-verbal communication [Goodwin 1980]. In a conversational setting, the head and eyes act as moderators: indicating thought, attentiveness, comprehension, engagement, and turn transitions to mediate the flow of conversation [Cassell et al. 1999]. While hand gestures and postural shifts also support communication [Cassell et al. 2001], the role of head and eye motion as non-verbal cues cannot be understated. Conversational head and eye animation is a complex interplay of personality, culture, psycho-linguistics, and scene context [Rossano 2012]. We present S^3, an animator-centric solution to generating such head and eye motion from input *speech* audio, a Director tagged *script*, and a cinematographic 3D *scene* (Figure 1).

Audio-driven facial animation research, and commercial solutions [Edwards et al. 2016; Karras et al. 2017; Richard et al. 2021; Zhou et al. 2018] have predominantly focused on the verbal production

of speech by the lower face. While audio correlations [Karras et al. 2017], or paralingual heuristics [Edwards et al. 2020a] can animate the upper face, head+eye rotations are left to be animated with the rest of the articulated body. As a result, most synthetic talking faces look straight ahead, despite psycho-linguistic research stressing that ≈ 30% of a conversation can be spent looking away from an interlocutor [Dawson 2022]. S^3 addresses this problem.

Conversation driven gaze has been used effectively since the birth of film to guide an immersive narrative, drawing audiences into the camera frame [Osipa 2010], and research on its psychological underpinnings date back half a century [Kendon 1967]. Prior art on speech-driven conversational gaze has typically been based on procedural psycho-linguistic heuristics, or data-driven models trained without a cinematographic scene context [Ruhland et al. 2015]. While recent research has addressed non-conversational gaze synthesis based on scene context [Goudé et al. 2023], conversational gaze without scene context [Jin et al. 2019], or viewer gaze [Boccignone et al. 2020], we are arguably the first to present a comprehensive, animator-friendly model of head and eye rotation in a conversational scene (see Table 1).

Our key insight, distilled from animation practice [Osipa 2010], is that while speech audio is primarily responsible for the pattern and timing of gaze aversion from a conversational partner, the precise 3D location of this gaze focus/aversion is largely determined by the cinematographic scene context. Our solution S^3 exploits this observation, to judiciously break down conversational head and eye motion into a number of animator-friendly components: *speech* audio-driven rhythmic head motion (eg. head nods) and transitions of focus/aversion of gaze from a conversational partner; *script*-driven emblematic head and eye gestures; and *scene*-driven saliency to contextually refine gaze focus/aversion into a temporal sequence of 3D look-at points (gaze trajectories), that our gaze control algorithm satisfies with optimal head and eye rotations.

Contribution: Our principal contribution is thus the design of a novel end-to-end solution for conversational gaze control S^3, built on the idea of audio-driven gaze focus/aversion and scene-driven 3D gaze refinement. Integrated into a typical animation pipeline in *Maya*, S^3 automatically computes head and eye rotations, relying on a speech-centric solution like JALI [2016] to animate remainder of the face. Further advances by S^3 include: a diarised and annotated dataset of conversation audio and inferred 3D scene context (with audio-visual processing code); an audio-driven neural model for rhythmic head motion; an audio-driven neural model that predicts temporal transitions of gaze focus/aversion from a conversational partner, which are refined by a 3D scene context to produce gaze trajectories; a gaze control algorithm, that generates head and eye animation to optimally satisfy given gaze trajectories.

Overview: A review of related work on audio-driven and head+eye animation (Section 2), is followed by terminology and a formal problem statement (Section 3). Section 4 presents data preparation: the selection of conversational videos, and the audio-visual processing to diarise, annotate and compute head+eye orientation from the videos. Section 5 details the algorithms for each component including: audio-driven rhythmic head rotation; audio-driven transitions in conversational gaze; 3D look-at point planning; and head+eye rotations to satisfy the sequence of 3D look-at points. We comprehensively evaluate S^3 in Section 6. Quantitatively, we validate S^3 using ground truth data, and compare it against baseline alternatives. Qualitatively, we present a large number of compelling and varied animations of conversational gaze; we show examples of directorial control over the resulting gaze; and animator critique of our workflow and results. A perceptual study shows S^3 to favor comparably against prior approaches to conversational gaze. Section 7 concludes with limitations and a discussion of future work.

2 RELATED WORK

Conversational head and eye animation lies at the intersection of speech audio-driven animation and head+eye animation, and is strongly informed by research of psycho-linguistic behavior. Ruhland et al. [2015] present an excellent review of inter-disciplinary research on gaze relevant to facial animation.

2.1 Psycho-linguistic head and eye motion

A rich body of literature going back more than 50 years has documented the role of the head and eye in non-verbal communication during a conversation [Argyle and Dean 1965][Argyle and Cook 1976]. Gaze transitions [Kendon 1967], have at least three communicative functions. *Turn-taking to mediate dialogue*: averting gaze when starting to speak, and looking back at the listener to conclude a turn [Ho et al. 2015][Bavelas et al. 2002]. *Monitoring understanding*: using gaze for lip-reading to better comprehend speech [Lusk and Mitchel 2016], or looking at the upper face to understand emotion [Buchan et al. 2007]. *Managing arousal*: looking away during moments of heightened emotion, high cognitive load [Doherty-Sneddon et al. 2012; Glenberg et al. 1998], social anxiety [Weeks et al. 2013], or when speaking with someone in power [Acarturk et al. 2021].

Gaze can also be consciously used as gestures (eg. elevator eyes, eye rolls) or for deictic purposes [Morency et al. 2006]. Gaze is further attracted by visual stimuli [Yoo et al. 2021], and people with status [Foulsham et al. 2010]. Cultural norms also impact head and gaze motion. For example, South Asians shake their head to agree, Arabs and Asians engage in mutual gaze more than Americans, and Chinese tend to look up while Japanese speakers look down when thinking [Haensel et al. 2022; Jan et al. 2007; McCarthy et al. 2008]. A number of multi-modal systems have been informed by these observations in combining face and body motion to create virtual characters with personalities [Cig et al. 2010; Sonlu et al. 2021].

We model all gaze behavior that is not directly related to *speech*, or visual stimuli in the *scene* using tags in a Directorial *script*.

2.2 Speech-driven animation

While research on computer facial animation dates back half a century [Parke 1998], there has been a recent surge of interest in digital humans in general, and speech-driven animation in particular. A large body of work spanning 25 years [Bregler et al. 1997; Thies et al. 2020], is video-based. In 3D, techniques can be classified as *procedural* (eg. [Edwards et al. 2016, 2020a]), *data-driven* (eg. [Karras et al. 2017; Richard et al. 2021]), or driven audio-visually by *performance capture* (eg. Face-off [Choi et al. 2022; Weise et al. 2009]). We

refer the reader to recent papers [Fan et al. 2022; Pan et al. 2022] for a more comprehensive survey of audio-driven lip-sync animation. Audio-driven head, hand and body gesture output have also been explored [Ghorbani et al. 2023; Kucherenko et al. 2020; Marsella et al. 2013; Stone et al. 2004].

In the context of audio-driven head+eye animation, the speech audio provides a tempo for rhythmic head motion and important psycho-linguistic cues for gaze focus and aversion.

2.3 Head motion

The head of a speaker or listener, is never perfectly still in conversation, constantly communicating through rhythmic and emblematic co-speech gestures, the absence of which make a character seem robotic [Frampton-Clerk and Oyekoya 2022]. A number of approaches to generating co-speech head and body gestures have been explored [Ghorbani et al. 2023]. State machines [Cig et al. 2010] and Hidden Markov Models [Zoric et al. 2011] to select between a set of head gestures such as a head nod or shake, based on prosody, using arousal, and dominance to head velocity and head direction, are over a decade old. Recent work such as Gesticulator [Ao et al. 2022; Kucherenko et al. 2020] and many submissions to the GENEA challenge [Yoon et al. 2022] train deep learning models to produce skeletal upper body animation from audio. Various image-based talking face methods [Biswas et al. 2021][Wang et al. 2021] also explicitly learn overall (rhythmic, emblematic and gazed based) head motion rendered together with an animated face.

We are inspired by these approaches to learning head motion from audio, but must additionally dis-entangle rhythmic head motion from head motion caused by controllable gaze transitions.

2.4 Gaze trajectories

Human gaze has been extensively studied in Robotics, Computer Vision, HCI and Graphics [Ruhland et al. 2015]. Most works delve into the dynamics of specific types of eye movements, such as micro-saccade and pupil-dilation [Duchowski et al. 2015], gaze shifts [Young and Stark 1963], and smooth pursuit [Meyer et al. 1985]. Patterns of gaze as attentive behavior using visual salience have also been studied [Cerf et al. 2007; Itti 2006; Marat et al. 2009], recently using face detection to amplify the salience on (speaking) human faces [Boccignone et al. 2020; Shi et al. 2020; Sugano et al. 2013]. Networks to predict gaze trajectories from input video and motion capture [Kerkouri and Chetouani 2021; Klein et al. 2019], and HMMs to synthesize gaze shifts between regions of a segmented face [Duchowski et al. 2019] have been studied. These methods however, credibly model the gaze of an observer and not the gaze behavior of a speaker.

There is less work on inferring the gaze behavior of characters in conversation. Procedural models directly encode psycho-linguistic heuristics [Ruhland et al. 2015], such as tagging audio for silences, and making the eyes look up at those times to simulate thinking [Zoric et al. 2011]. We use tagged scripts similarly, but only for emblematic gestures, cultural preferences, and behavior that cannot be automatically inferred from speech audio, and a 3D scene context. Prosody and previous frames of video have been used to train a linear binary classifier to forecast gaze focus/aversion [Ward et al. 2016]. Speech audio agnostic to a 3D scene, has been used to determine eye and head motion using rules [Marsella et al. 2013], or learnt from data [Jin et al. 2019; Le et al. 2012], Real-time learning of eye motion alone, given a user's audio and head motion as input, has also been attempted, for VR applications [Canales et al. 2023].

Conversely, non-conversational gaze models based on 3D scene salience, inhibition (to prevent persistent gaze fixation) [Goudé et al. 2023; Pan et al. 2020], and the body motion of a character interacting with the scene [Pejsa et al. 2016] have also been explored.

In contrast, we present a comprehensive model for conversational head+eye motion. Motivated by animator workflows, we combine audio-driven ego-centric gaze focus/aversion, refined by exo-centric 3D scene context, to compute a sequence of 3D gaze transitions.

Gaze Control. Given a gaze trajectory (i.e. a temporal sequence of 3D look-at points), inverse computing head (2 DOF) and eye (2 DOF) rotations to satisfy the look-at points is an under-constrained problem, and typically involves both a head and eye rotation [Leigh and Zee 2006]. Proximal gaze targets ($\approx< 20^o$) can be achieved by rapid eye-only gaze shifts called *saccades*, with well-studied velocity profiles. The relative timing and amount of head and eye motion can vary based on the gaze shift needed for the target, the time to target, whether the target point is pre-planned or reactive, and the intended dwell time on the target [Leigh and Zee 2006]. A common approach to this problem in graphics is to use an eye-only gaze shift threshold beyond which both eye and head rotate [Normoyle et al. 2013]. Other approaches include mass-spring models of *smooth pursuit* dynamics [Itti et al. 2004], and combinations of saccades and smooth pursuit [Yeo et al. 2012]. A parametric model providing control over many aspects of a desired gaze target is useful in a variety of behavioral contexts [Andrist et al. 2012]. Data-driven [Jin et al. 2019] and emotionally expressive [Ferstl 2023] gaze control have also been explored.

Our model draws inspiration from these approaches and computes head and eye rotations, as an optimization that includes a novel term that accounts for the dwell time of a look-at point.

Table 1. Research on automated prediction of gaze behavior

Research work	Audio input	Scene input	Rhythmic head output
[Jin et al. 2019]	✓	×	×
[Le et al. 2012]	✓	×	✓
[Canales et al. 2023]	✓	×	×
[Pan et al. 2020]	×	✓	✓
[Goudé et al. 2023]	×	✓	✓
S^3	✓	✓	✓

3 PROBLEM STATEMENT

We now formally state our problem of conversational gaze animation in terms of inputs: **speech**, **script** and **scene**; and outputs: **head** and **eye** animation.

Speech. audio signal comprises audio streams $A_1(t)$ and $A_2(t)$ for two speakers in dyadic conversation, where time $t \in \{1..T\}$ is

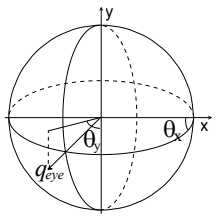
T frames of animation. While we describe a two-person conversation through-out, our approach applies seamlessly to n speaker conversations, where the interaction is dyadic (Section 5.5). A single audio stream input (see Video), is readily diarized into two or more streams using NeMo [2021]. We automatically obtain an aligned speech transcript from audio using [Radford et al. 2022a].

Script. provides animators with an *optional* Directorial interface to script head+eye behavior, trigger emotion/emblematic gestures etc. Tags of the form <start><end/> are embedded within the audio-aligned speech transcript [Radford et al. 2022a]. The tags are extendable, with present support to spatially modulate the scene salience such as <avert-up><avert-up/> to author a preferred direction of gaze aversion, and overrides that can force certain gaze behavior.

Scene. provides *optional* spatio-temporal information about visually salient parts of the conversational setting. We model the scene using 3D positions p_1, p_2, neutral facing directions $\vec{d}_1, \vec{d}_2$ for two speakers, and the 3D position $\{v_i(t)\}$ and saliency weight $\{s_i(t)\}$ of k animated visual hotspots $i \in \{1..k\}$. Such hot-spots are easily authored in a 3D scene, can be inferred automatically, or derived from intensity maps of visual saliency [Goudé et al. 2023].

Head. is modeled local to the neck/body transform B as a 3DOF rotation vector θ_h with pitch, yaw and roll as rotations about x, y, z respectively, that define a local head transform H. We ignore the contribution of head roll (z axis rotation) in controlling gaze.

Eyes. are modeled using a 3D world space look-at point q. For an eye at point e local to the head, $q_{eye} = (BH)^{-1}q - e$. The 2DOF pitch and yaw x, y rotation vector θ_e for the eye are the spherical polar co-ordinate angles of q_{eye} as shown in the inset. Representing an eye as a world space look-at point has advantages: most animator rigs use a global look-at point as an eye rotation controller, aligned with an oculocentric motor strategy [Henriques et al. 2002]; and the Vestibulo-Ocular Reflex movement is inherently captured [van der Steen 2009].



4 DATA PREPARATION

Many parts of our solution S^3, such as the audio-based aversion prediction model (section 5.2) and rhythmic head model (section 5.5), consist of deep learning components. We considered a number of audio-visual datasets, but they either did not capture diarised and annotated dyadic conversations in a sparse natural setting, or were too small to support deep learning. Notable among these was the 17 minute long Cardiff conversation dataset [Vandeventer et al. 2015] (our dataset in contrast is 379 minutes long). We thus curated a new *audition* dataset. Here, we discuss our data sourcing, and the audio-visual processing used to annotate its head and eye motion, gaze fixations, and the decoupling of rhythmic co-speech gestural head motion from head motion due to gaze shifts.

4.1 Dataset Source

We sourced our data from in-the-wild acting audition performances found on *Youtube* (similar to other publicly sourced datasets [Ephrat et al. 2018; Nakazawa et al. 2020]). The videos all have one on-screen actor, and one off-screen actor, engaging in a conversation. We chose these videos for two reasons: one, unlike TV interviews and talk shows, which often cut from speaker to speaker, the actor being auditioned is always in the frame in an audition clip, providing data and insight for both speaking and listening behaviors; and two, actors are less inhibited by a camera and their performances tend to be varied, natural, and expressive, compared to those captured in a lab setting.

Our *audition* dataset is comprised of 111 audition videos (a list is provided in supplementary material) with a total length of 379 minutes. Overall in the videos, the on-screen actor spends approximately 63% time speaking, and 37% time listening (where the off-screen actor is speaking).

4.2 Head+Eye Gaze Annotation

We annotate each video frame using binary labels. Each video frame is labelled as either "gaze-on", "focused" (0) when the on-screen actor is looking at the off-screen actor, or "gaze-off", "averted" (1) when their gaze is directed elsewhere. The labelling is used to train our audio-driven gaze aversion probability network (section 5.2). We use an off-the-shelf gaze-estimation model [Zhang et al. 2020], to obtain gaze direction from the video (Figure 2). We then use a dispersion-based filtering technique [Birawo and Kasprowski 2022] to ignore micro-saccades, reduce jitter, and segment the gaze signal into a sequence of some N gaze fixations, with direction $\vec{p}_i$ over time interval $< t_s, t_e >_i$, where $i \in \{1..N\}$. Based on the insight that speakers in an audition tend to spend the majority of the time looking at the conversation partner, we use a Gaussian mixture model to cluster $\vec{p}_i$, and use the center of the biggest cluster as the direction $\vec{p}_{off}$ towards the center of the off-screen actor. We represent the angular size of the off-screen actor as a cone angle ϕ ($\phi \in [0, \pi/2]$) around $\vec{p}_{off}$. A gaze direction $\vec{p}$ is thus averted from the off-screen actor iff it deviates $>= \phi$ from $\vec{p}_{off}$. For unit gaze vectors:

$$averted(\vec{p}, \vec{p}_{off}, \phi) = \lceil cos(\phi) - (\vec{p}.\vec{p}_{off}) \rceil \quad (1)$$

Given that dispersion-filtering removes micro-saccades, we would like the majority of the remaining gaze shifts to and from the off-screen actor to count as gaze focus/aversion transitions. We thus do a line search on $\phi \in [\epsilon, \pi/2]$ to maximize the total number of focus/aversion gaze transitions, where ϵ provides a minimum speaker size angle (we pick ϵ as the smallest cone angle to contain half the gaze directions in the off-screen actor cluster). In other words:

$$\max_{\phi}\left(\sum_{i=1}^{N-1} |averted(\vec{p}_i, \vec{p}_{off}, \phi) - averted(\vec{p}_{i+1}, \vec{p}_{off}, \phi)|\right) \quad (2)$$

Finally, we use *averted* to label video frames, and our results strongly match viewer expectations (see Video 1:58-2:09).

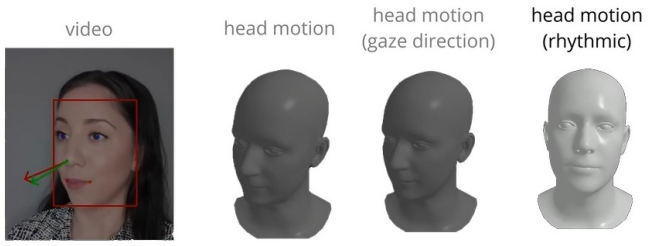


Fig. 2. *Audition* data with head (red) and gaze (green) estimation (left), and isolated rhythmic head rotation (right).

4.3 Extracting Rhythmic Head Motion

In order to train a model for predicting rhythmic head motion using audio and text transcript features, we need to isolate rhythmic head movements in our *audition* dataset from gaze-driven head motion. To identify eye-driven head movements, we implemented a Dynamic Time Warping (DTW) based algorithm [Giorgino 2009]. Note that DTW is necessary because while the head always moves complementary to the eyes, it is often delayed (100-200ms) and always moves slower [Ruhland et al. 2015]. Our DTW measures the optimal time-warped similarity between the temporal rotations of gaze $\theta_{eye}(t)$ and head $\theta_{head}(t)$ ($\theta^z_{head}(t)$ is ignored in the comparison).

We first determine gaze and head rotations for our dataset. We use the ETH-XGaze model [2020] to compute eye rotations $\theta^{x,y}_{eye}$, and Mediapipe [Lugaresi et al. 2019] for head rotation $\theta^{x,y,z}_{head}$, from the videos. Both head and eye rotations are de-noised using a Gaussian filter. These are then given as input to the DTW algorithm which first determines L_2 distance $d(e_i, h_j)$, between each pair of frames e_i in $\theta_{eye}(t_s)$ and h_j in $\theta_{head}(t_s)$, where t_s indicates the sliding window samples from the eye and head rotation sequences [Kucherenko et al. 2020](Figure 11). A cost matrix, C, of size $n \times m$, is constructed where n is the length of $\theta_{eye}(t_s)$ and m is the length of $\theta_{head}(t_s)$. The cost matrix cells (initialized to ∞), are iteratively filled to compute the minimal cost based on neighboring cells:
$C(i, j) = d(e_i, h_j) + min(C(i-1, j), C(i, j-1), C(i-1, j-1))$.
We accumulate the dissimilarity along different possible paths, in an accumulated cost matrix D as:
$D(i, j) = C(i, j) + min(D(i-1, j), D(i, j-1), D(i-1, j-1))$.
Starting from $D(n, m)$, we backtrack through D, to find the optimal warping path (left, diagonal, or up at each step) ending at $D(1, 1)$, with the smallest accumulated alignment cost $D_{optimal}$. The rhythmic head movement is then calculated as follows: For head rotation samples with a low alignment cost ($D_{optimal} \leq \tau$, where τ is the mean of all optimal alignment costs for the entire video), head and gaze are correlated; we subtract the aligned gaze rotation from the head rotation sample to get the head rotation sample $h_l(t_s)$. For head rotation samples with a high alignment cost ($D_{optimal} > \tau$), head and gaze are independent, and we orient the mean pose of the sample to the front-facing rest head pose, and create a new head rotation sample $h_h(t_s)$. Finally, we concatenate the rhythmic head rotation samples $h_l(t_s)$ and $h_h(t_s)$ as originally aligned in time and use interpolation to remove any remaining discontinuities due to shot changes, noise in head/eye tracking, extreme face rotations and occlusions, to produce a rhythmic head motion signal $\Delta\theta_{head}(t)$ (see Video 2:14-2:31).

5 S^3 ALGORITHM

We now present algorithmic details for animating dyadic conversational gaze in S^3 (Sections 5.1-5.7), and its extension to N-party conversations (Sections 5.8).

5.1 Algorithm Overview

S^3 takes in three streams of inputs: including speech audio, 3D scene context, and optional scripts. As outputs, head and gaze trajectories are produced. Architecturally S^3 consists of three independent modules (Figure 3): A deep-learning-informed look-at-point (gaze trajectory) planner, an inverse kinematics (IK) gaze controller and a learned rhythmic head motion generator inspired by [Kucherenko et al. 2020]. As the first step, the look-at-point generator creates time sequences of gaze transition targets $\{t_i, \vec{q}_i\}_i^N$, for each character in the conversation. These gaze targets are then processed by our gaze control IK model, to create realistic per-frame trajectories of head rotation $\bar{\theta}^{x,y}_{head}(t)$ and gaze $q(t)$. Finally, we use the rhythmic head controller to produce rhythmic head motion $\Delta\theta^{x,y,z}_{head}(t)$, which is added to gaze-based head motion $\bar{\theta}_{head}(t)$ to generate the final output $\theta_{head}(t)$. As mentioned in Section 3, the *scene* context provides

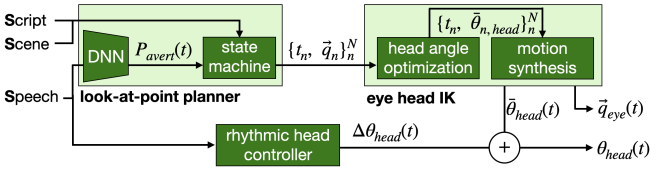


Fig. 3. Method Overview

3D positions p_1, p_2, and neutral facing directions $\vec{d}_1, \vec{d}_2$ for two speakers, as well as the 3D position $\{v_i(t)\}$ and saliency weight $\{s_i(t)\}$ of k animated visual hotspots (potential look-at points) $i \in \{1..k\}$. The *scene* context, along with *speech* audio, and optional *script* tags (Section 5.6), comprise the input to our algorithm (Figure 3). Our modular look-at-point planner further allows S^3 to easily handle three-party (or n-party) conversations (Section 5.8). This level of extensibility and control is hard to achieve without a modular design.

5.2 Look-at-point planner: aversion probability network

We predict a *speech* based probability $p_{avert}(t)$ for a conversational agent to avert their gaze from a conversational partner at every timestep t, using a recurrent neural network architecture (Figure 4).

Our network uses two forms of input: prosodic audio features encoded using Mel Frequency Cepstral Coefficient (MFCC), log filter bank energies, and Spectral Subband Centroids (SSC); and the relative timing of speaking/listening turns obtained from audio-aligned speech transcripts. These inputs are commonly used in speech-based classification tasks [Rahmeni et al. 2020]. Swapping the input speech streams X_0 and X_1 in our symmetric model, allows us to predict the gaze aversion probability of the conversational partner $P^1_{avert}(t)$.

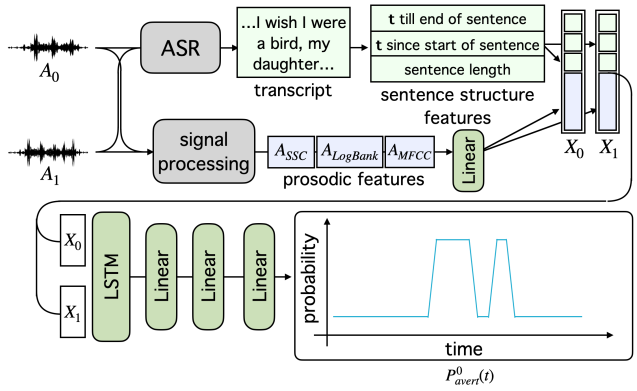


Fig. 4. Gaze Aversion Prediction (updated figure)

The model is trained on our *audition* dataset. After a 9:1 train-test split, we divide each audio performance into 10-second segments, with 5 seconds of overlap between them. The model is then trained with binary entropy loss to produce output that matches the aversion state (0 or 1) labeled in section 4.2. Model parameters are updated using the Adam optimizer with default parameters [Kingma and Ba 2014], training stopped after 1400 epochs. Our model **achieves 98.4% and 78.9% accuracy on training and validation sets** respectively, and generates gaze aversion probabilities that are overall smooth. See Section 6.1 for a more comprehensive evaluation.

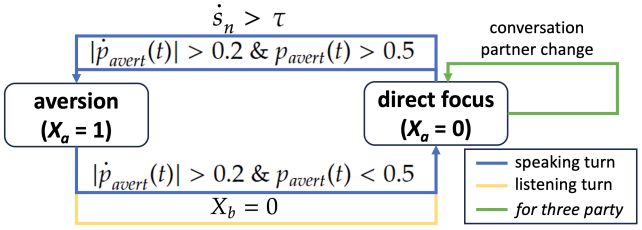


Fig. 5. Per-frame gaze state machine, green arrow is relevant in section 5.8

5.3 Look-at-point planner: state machine

For each conversational agent a, the look-at-point planner operates an aversion state machine $X_a \in \{0, 1\}$, switching between direct focus (gaze-on =0) and aversion (gaze-off =1) states every time-step. Direct focus generates look-at-points on the conversation partner, and aversion employs a random walk algorithm to generate look-at-points based on scene salience. The state machine transition is informed by three inputs:

-The speech-based gaze aversion probability $p_{avert}(t)$ (Section 5.2).
-The visual salience of each scene object $s_n(t)$.
-The human tendency to mutually engage gaze [Goodwin 1980], using the gaze state of the conversational partner $X_b(t)$.

As shown in Figure 5, a change of gaze state X_a at time t, is primarily controlled by the speech-driven probability of gaze aversion $p_{avert}(t)$, but can also be triggered by attending to a scene object n with a large increase in salience $s_n(t)$, i.e. $\dot{s}_n(t) > \tau$ (default $\tau = 0.5$ for saliency $s_n \in [0, 1]$). As the *audition* videos are visually focused on one agent, mutual gaze is not explicitly captured by the learnt gaze probability $p_{avert}(t)$. We can model mutual gaze by coupling the the state machines of the conversational agents, so that an averted agent ($X_a = 1$) can transition to $X_a = 0$ to match direct gaze from the conversation partner $X_b = 0$. We handle the coupled state machines of agents a and b in two passes. In the first pass, we generate the gaze states of both agents a and b on their speaking turns, considering only the signals $p_{avert}(t)$ and $\dot{s}_n(t)$. In the second pass, we generate gaze states for the listening turns of both agents, using $p_{avert}(t)$, $\dot{s}_n(t)$, and X_b (or X_a) computed for the speaker in the first pass (see Video 4:42-4:57 for example).

Once the gaze of both conversing agents has been classified as direct or averted for each frame, we compute a time sequence of gaze fixations. Deviation from the fixations are modeled as microsaccades (Section 5.7). The fixated look-at-points are computed as follows:

Direct Focus: The agent looks at the other interlocutor (center of the face by default).

Averted: We employ a random walk similar to [Boccignone et al. 2020] to generate a sequence of scene salient look-at-points. The duration of each look-at is sampled from a known distribution of human fixation [Goudé et al. 2023], and the choice look-at-point sampled from a weighted distribution that favours object salience, and gaze shifts of small amplitude. Specifically, when selecting a new gaze target, we compute ρ_i for scene objects $i \in \{1..k\}$ as:

$$\rho_i = s_i \cdot e^{-\kappa \cdot \max(1, 1/dur) \cdot \|v_i - v_{prev}\|}$$

where $\kappa = 1.33$ based on [Boccignone et al. 2020], s_i and v_i are the salience and position of the i^{th} object at the current time, v_{prev} is the previous look-at point, and dur is the length of the aversion interval. We then use the soft-max function to compute a probability distribution from ρ_i, from which we select the new scene object (look-at-point). Different from [Boccignone et al. 2020], we also use the aversion duration dur, to ensure a small gaze shift for a very short ($< 1sec$) gaze aversions. Finally, we sample the time of the next gaze shift from a distribution of fixation duration (shifted gamma law with $\alpha = 1.2394$, $\theta = 0.1880$, and $loc = 0.08$) [Goudé et al. 2023].

5.4 Gaze Control IK

Our gaze control IK algorithm augments prior art solutions [Jin et al. 2019; Yeo et al. 2012], to present improved generation of head+eye motion for our context, given a time sequence of gaze targets. We first solve an optimization problem for the head contribution to each gaze shift; then, use a motion generator to interpolate the desired sequence of head and eye targets.

Head Contribution Optimization: Given look-at-point planner gaze targets, we determine the required head rotation as an optimization of three terms to: match a learnt co-relation between head and gaze angles [Jin et al. 2019]; minimize head rotation from its predominant focus on the other interlocutor; and minimize eye rotation needed to meet the gaze target. In other words:

$$\begin{aligned}\bar{\theta}_{head} = {}^{*}\, argmin_{\theta}(& w_p * \|\theta - \theta_p\|^2 \\ & + w_n * (1 - dwell) * \|\theta - \theta_n\|^2 \\ & + w_e * dwell * \|\theta - \theta_{eye}\|^2)\end{aligned}$$

where $\{w_p, w_n, w_e\}$ are constants weighting the three terms; $\theta_p = g(\theta_{eye})$ is a learned mapping of the most probable head angle, for a given gaze direction, proposed in [Jin et al. 2019]; θ_{eye} is the direction that the gaze target makes with the neutral eye direction; θ_n is the direction facing the conversational partner, typically close to the neutral head direction; and $dwell = min(dur, 1)$ is a weight increasing with gaze target fixation time dur (clamped at 1).

Small $dwell$ penalizes head movement from neutral, encouraging eye motion to match the gaze target, and the opposite for large $dwell$, We determine the weights for each term using a grid search on different combinations of $\{w_p, w_n, w_e\}$ to find a set of weights that minimizes the Mean Square Error (MSE) with the annotated head and eye angles generated from our *audition* dataset (Section 4.2). Our proposed optimization results in a lower MSE **10.92** compared to **24.26** using $\theta_{head} = \theta_{eye}$ [Yeo et al. 2012], **16.04** using $\theta_{head} = \theta_n$, or **11.30** using $\theta_{head} = \theta_p$ [Jin et al. 2019].

Motion Generator: We use a modified version of the head-eye motion generator in [Yeo et al. 2012] to interpolate the sequence of target head and eye angles. For both eye and head motion, movement $\dot{\theta}(t)$ is produced by summing up a sequence of sub-movements:

$$\dot{\theta}(t) = \sum_{i}^{N} \mathbf{b}_i v\left(t_i^0,\ t_i^1,\ t\right)$$

where each sub-movement has a direction $\mathbf{b}_i$ and a velocity profile:

$$v\left(t^0,\ t^f,\ t\right) = \frac{30}{\left(t^f - t^0\right)^5}\left(t - t^f\right)^2\left(t - t^0\right)^2$$

The velocity profile for head and eye sub-movements differ by motion duration (100ms for the eye, and 600ms for the head). [Yeo et al. 2012] breaks down a large gaze shift into a sequence of smaller saccades that look more realistic. Every 200ms, an eye sub-movement $\mathbf{b}_i$ is generated towards a position predicted by their character's imperfect probabilistic perception model. We achieve a similar effect by artificially adding noise to the specified look-at-point θ_{eye}:

$$\begin{aligned}
\dot{\theta}_{target,\mu} &= \alpha(\theta_{eye} - \theta_{prev}) \\
\dot{\theta}_{target} &\sim \mathcal{N}\left(\mu = \dot{\theta}_{target,\mu},\ \sigma = \tfrac{1}{4}(1-\alpha)||\theta_{target,\mu}||\right) \\
\theta_{target} &= \theta_{prev} + \dot{\theta}_{target}
\end{aligned}$$

By ensuring $\alpha > 0.5$, we guarantee that each gaze shift gets closer to the target look-at-point. Once the current look-at-point is sufficiently close to the target, we simply use $\theta_{target} = \theta_{eye}$ to prevent oscillation about the look-at-point. We use a similar strategy for head sub-movements, except we set $\sigma = 0$ to ensure smooth head motion.

5.5 Rhythmic Head Controller

We train an audio-driven neural network to generate *rhythmic head motion* (section 4.3). We use an architecture inspired by *Gesticulator* [2020] to produce rhythmic head rotation values at every time-step, shown in Figure 6. Audio and textual features serve as inputs. For audio, we use Mel-spectrogram along with the prosody information (intensity and pitch) of the audio. For text, we use Bert features [Devlin et al. 2019] and the sentence structure features outlined in Figure 4.

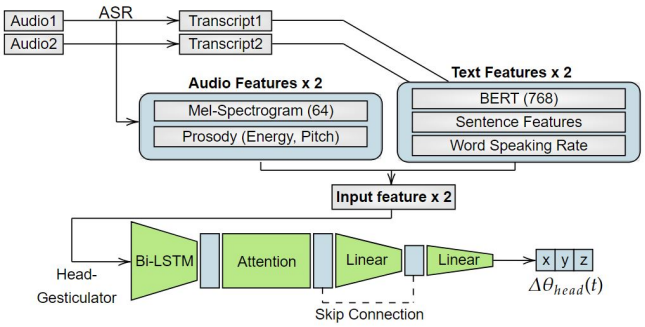


Fig. 6. Rhythmic Head Gesture Model

We train on the *audition* dataset for 100 epochs using weighted MSE loss for both velocity and position (we weigh samples further away from the mean at a higher weight). We verified that our model predicts dynamic motion instead of a static mean by observing that the position and velocity distribution generated by our model closely resembles that of the dataset. (Figure 7).

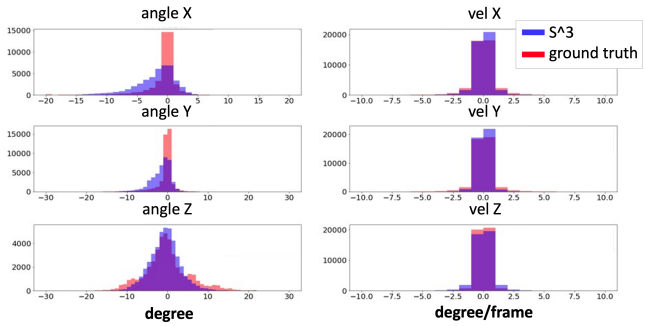


Fig. 7. Rhythmic Head Motion Prediction

5.6 Animator Scripting

Our modular architecture supports animator control in various places. Speech transcripts embedded with directorial tags are a popular approach to controlling animation in games [Edwards et al. 2020b]. We thus demonstrate animator control in S^3 using extendable tags embedded in a [Radford et al. 2022b] generated speech text transcript.

We support three kinds of tags. *look-at* tags, which amplify the salience of an object while the tag is active, causing an agent to focus on an important object, or reflect specific gaze behavior, like looking out a windshield while driving. *directional* tags are used to specify ego-centric aversion behavior such as averting up to reflect thinking, or averting down due to reflect guilt. Such tags zero out the salience of scene objects in the opposite direction for the duration of the tag. *override* tags can force focus/aversion labelling over the tag's duration, for example to specify speech agnostic concentration.

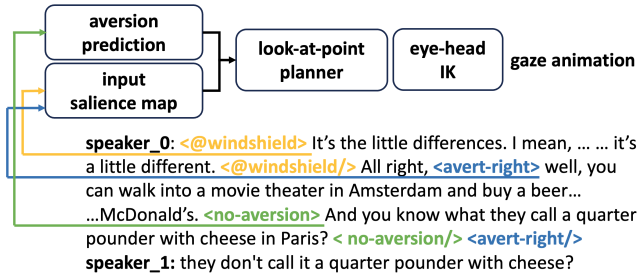


Fig. 8. Tag varieties and their control: look-at tags (yellow); directional tags (blue); gaze-on/gaze-off tags (green)

5.7 Microsaccades

When fixated on an object, humans perform small (<2 degrees) and frequent (1-2 Hz) saccades [Chung et al. 2015] within the object to prevent perceptual fading (where vision blurs due to de-sensitized neurons). Microsaccades are found to be essential to the realism of gaze animation [Krejtz et al. 2018]. We model microsaccades during fixation as a post-processing step similar to [Pan et al. 2020]. During any gaze fixation interval longer than 0.5 seconds, we sample irregular intervals from $\mathcal{N}(0.5, 0.1)$, where we induce a small eye rotation $\Delta\theta_t$ of amplitude $\mathcal{N}(0, 2)$, that is added to the output gaze animation to enhance realism.

5.8 Three-Party Conversation

Unlike end-to-end regression approaches like [Jin et al. 2019; Le et al. 2012], our modular approach can be readily adapted to N-party conversations with some simple modifications. We illustrate this extension to a three-party conversation with agents a, b, and c, where we assume people speak one-at-a-time. We cast this scenario as pairs of dyadic conversations. From the perspective of a, when b or c is speaking, it is a dyadic conversation between $a + b$ or $a + c$, respectively. When a is speaking, it is a dyadic conversation between a and the previously speaking agent. The third interlocutor in all cases is simply treated as a salient scene object.

We can thus dynamically re-register the conversation partner for each agent when speaking turns change, and re-use our dyadic algorithm. Further, changing a conversation partner automatically triggers a gaze shift. The 3-party conversation model is implemented in comparison to [Jin et al. 2019] in section 6.2.

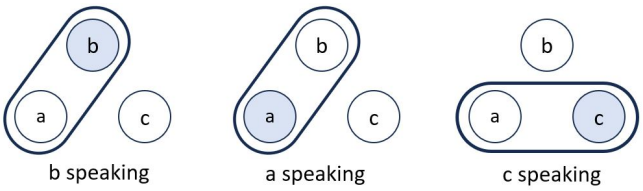


Fig. 9. 3-party conversations cast as two dyadic conversations involving $a + b$ and $a + c$

6 EVALUATION

We manually checked about 10% of the *audition* videos to confirm that both, our gaze annotation, and rhythmic head motion computation (Section 4), strongly matched viewer expectation.

We also presented quantitative validation of each technical component of S^3 in Section 5. Specifically we report:

- 98.4% and 78.9% accuracy on training and validation data, for our aversion probability network (Section 5.2).
- Our state machine when correctly averted (Section 5.3), picks the correct aversion gaze cluster in the *audition* dataset (Section 4.2) with 90.7% accuracy.
- Our predicted IK head angle for gaze fixations (Section 5.4) has a lower Mean Square Error of 10.92° (compared against the *audition* dataset) than prior art. While fixated head+eye values in the *audition* dataset are reliable, their motion trajectories can be noisy, and thus we do not compare it to our head+eye motion interpolation ouptut.
- Our rhythmic head controller produces a distribution of rhythmic head motion (Figure 11) that closely matches the *audition* dataset (Section 5.5).
- We show effective animator control using a variety of script tags (Section 5.6).
- We show S^3 can be easily adapted to generate gaze for pairwise dyadic, N-party conversations (Section 5.8).

We further present a comprehensive analysis of our aversion prediction network, the core of S^3. This is followed by a perceptual study comparing our output with prior art [Jin et al. 2019; Le et al. 2012]. We present a large number of compelling and varied animations of conversational gaze, and animator critique of our workflow and results.

6.1 Aversion Prediction Network Evaluation

Beyond high per-frame accuracy in predicting a gaze focus/aversion state, we analyze the performance of our network on various metrics, compared to a few baselines. Specifically we compare against *stare* a commonly used model with no gaze aversion; and a *statistical* model that alternately samples gaze focus/aversion aversion intervals randomly, from distributions of focus/aversion interval length in the *audition* dataset. The outputs of the three models, relative to ground truth, for an example 20 second clip are shown in Figure 10.

We evaluated each model's predictions $\{\hat{p}_n\}_{n=1}^N$ against ground truth data $\{p_n\}_{n=1}^N$ using accuracy, Jaccard similarity (IOU), gaze-on/off transition accuracy, and aversion instance ratio.

Accuracy measures the per-frame agreement between $\hat{p}_n$ and p_n i.e. $(acc = 1 - (\sum_{n=1}^N |\hat{p}_n = p_n|)/N)$. Jaccard similarity measures the frame overlap between predicted gaze aversion and ground truth ($\mathbb{1}$ is the indicator function below):

$$\sum_{n=1}^{N} \frac{\mathbb{1}\{\hat{p}_n = p_n = 1\}}{\mathbb{1}\{\hat{p}_| n = 1\} + \mathbb{1}\{p_n = 1\}}$$

Gaze-on (or off) accuracy is a binary measure of alignment between a predicted gaze transition and the closest ground truth, at perceptually significant moments of gaze transition from aversion to focus.

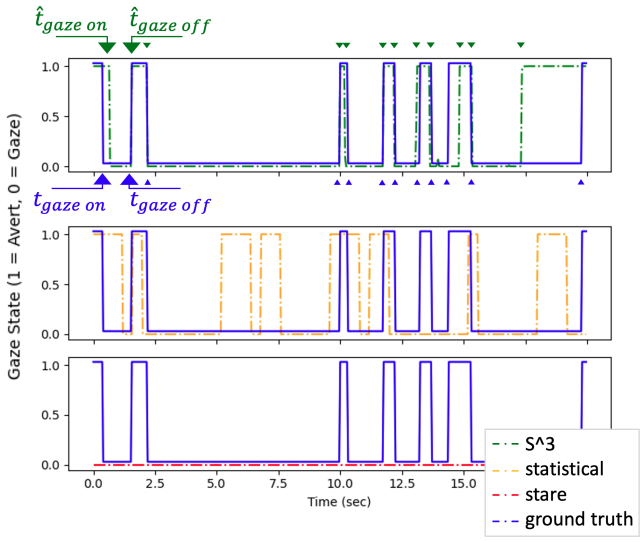


Fig. 10. Gaze focus/aversion using S^3, statistical, and stare vs. ground-truth.

The aversion instance ratio simply counts the number of aversions, relative to those in ground truth.

Table 2. Comparison of models to predict gaze aversion

Model	Acc	IOU	Gaze-on Acc	Gaze-off Acc	Avert Instances
Stare	0.63	0.00	0.00	0.00	0.00
Statistical	0.47	0.23	0.31	0.33	**1.04**
S^3	**0.79**	**0.36**	**0.53**	**0.53**	1.08

From the table, we can see that while stare (performing no aversion) achieves 63% accuracy (because gaze focus is predominant), it performs poorly on the other perceptual metrics. The statistical model also fails to generate gaze aversion at times that perceptually make sense. S^3 performs well on all metrics with high accuracy, Jaccard similarity, good alignment of gaze transition, and generates a similar number of gaze transitions as the ground truth.

6.2 Perceptual Prior Art Comparison Study

As far as we know we are the first to combine speech audio and scene context in a model for conversational head+eye motion (Table 1). the closest conversational gaze prior art to us are perhaps [Jin et al. 2019; Le et al. 2012]. We were unable to access code, executable or animation curves for either work. We thus ran our model using the audio from their their supplementary video examples (three from [Le et al. 2012], and two from [Jin et al. 2019]) (see Video 2:55-3:24 and supplemental).

We chose camera views and framing to match their output for our results. We then conducted a 4 point (weak or strong preference) forced choice user study with 36 users, between our output and [Le et al. 2012] or [Jin et al. 2019] (approved by ethics protocol #38139). We instructed users to focus on head+eye motion and ignore rendered appearance and other factors. We also asked the users to provide reasons for their choice, and overall impression of the animations. We eliminated results from 5 users because their choices were statedly made on the rendered appearance or the quality of lip-sync. Our final user demographics consist of 6 facial animators, 6 non-facial animators, and 19 lay viewers. The results of the forced choice experiment are shown below. A binomial test evaluates the significance of the result, with a p-value displayed on the top of each bar graph. It can be seen that our model compared favourably against both [Le et al. 2012] and [Jin et al. 2019].

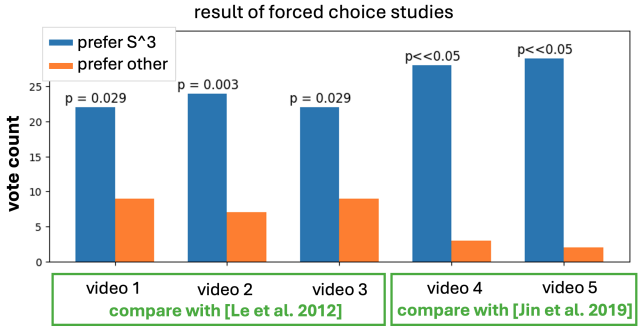


Fig. 11. Perceptual study comparing S^3 and [Jin et al. 2019; Le et al. 2012]

6.3 Animator and Casual Viewer Feedback

Facial animators noted that the head movements from [Le et al. 2012] were “too smoothed” but also “had many discontiniouties”, and that the movements look “repetitive”. Casual users found the gaze “aimless” and “does not look connected with speech”. They also noted “a lot of head movements” and we divided between it seeming "expressive" or "erratic".

Viewers felt [Jin et al. 2019] had very “static eyes”, head movement that looked “robotic”, and gaze that “lacked eye contact”.

In comparison, facial animators found S^3 to have "convincing mutual gaze", “reasonable gaze targets”, “high-quality motion control”, and generating “great aversion” that “fits the sentence structure” and audio. It was mentioned that S^3 eye movement was “a bit dramatic given the neutral manner and speaking content” of video 3. Casual users praised S^3 gaze as “sensible”, “natural”, and the performance as “lifelike”. On the critical front, some users felt the gaze was not as smooth as [Le et al. 2012], and that the S^3 "doesn’t generate enough head motion".

6.4 Cinematic Results

Finally, we include eight clips from film/TV (see supplemental video 8:38-17:34) from outside of our Audition dataset. Each clip was diarized, and a 3D scene with the speakers and 3-5 salient points created to match the clip. We additionally used scripting on "Royal with Cheese" and "Dear Dolores" to establish the contextual importance of a moving car windshield and reading a letter, respectively.

Fig. 12. Direct gaze and gaze aversion transition examples, taken from the Heat restaurant scene.

For each clip, we begin by creating audio-driven facial performances using JALI [Edwards et al. 2020a], on the character rigs shown for example in Figure 12. We then employ S^3 to automatically generate head and eye motion trajectories, that are mapped to control the head/neck and eye transforms on the rigs. Further, the animated head and eye rotations produced by S^3, can be easily combined with any existing head and eye motion to support a variety of rigs and workflows. The run-time for components of S^3 on a 1 min. audio clip (run on a RTX3060 GPU) are roughly: diarization 25s, audio2text 10s, audio features 30s; rhythmic head motion 2-3s, gaze planning and control 2-3s. We include additional automatically generated clips in supplementary material along with the comparison videos used in the perceptual study, showing nearly 15 minutes of conversational gaze animation.

7 CONCLUSION

While our evaluation shows S^3 to be an effective workflow to animating conversational gaze, our approach is not without limitations.

- perfect phonetic alignment between a Directorial script (transcript) and audio, remains a challenge for long audio clips with non-lexical or muffled sounds, noise, cross-talk and background chatter. While this poses a greater problem for lip-synchronization [Pan et al. 2022] (see inaccurate lip-sync around 15:16-15:18 and 16:15-16:19 of the supplementary video), it can cause problems with the timing of scripted gestures and other tags.

- the complexities of speakers talking simultaneously, cross-talk, multi-person group interaction, or speaking to a crowd however, are subject to future work.

- we do not exploit any emotional or cognitive information in the input speech audio or transcript, that could be extracted by sentiment analysis and used to modulate the output head and eye animation. Presently such information can be explicitly authored using a tagged script.

- we rely on existing audio-driven approaches to animate blinks and other paralingual behavior [Edwards et al. 2020a]. Conceptually this can produce undesirable results, as we do not explicitly model correlations between blinks and gaze transitions. In practice however, our blinks align well with the onset of gaze transitions.

In summary, S^3 is a novel modular approach to conversational head and eye animation. We model ego-centric gaze behavior as speech audio based transitions of gaze focus/aversion, refined by exo-centric gaze behavior based on 3D scene saliency, to output conversational gaze trajectories. A novel gaze control IK algorithm, then generates head and eye animation, to satisfy the conversational gaze trajectories, combined with audio-driven rhythmic head motion, and script-driven emblematic head+eye gestures. Favorable comparison to prior art, viewer critique, and compelling results show S^3 to be a 'sound' approach to audio-driven head and eye animation. We anticipate our insights and workflow to meaningfully impact audio-driven hand and posture animation, and inspire new directions in expressive facial animation.

ACKNOWLEDGMENTS

We would like to acknowledge Chris Landreth for his valuable inputs and artistic direction on this paper. Thanks are also due to Eugene Fiume, Pif Edwards, Sarah Watling, and NSERC Discovery Grant 480538.

REFERENCES

Cengiz Acarturk, Bipin Indurkya, Piotr Nawrocki, Bartlomiej Sniezynski, Mateusz Jarosz, and Kerem Alp Usal. 2021. Gaze Aversion in Conversational Settings: An Investigation Based on Mock Job Interview. *Journal of Eye Movement Research* 14, 1 (May 2021). https://doi.org/10.16910/jemr.14.1.1

Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. 2012. A Head-Eye Coordination Model for Animating Gaze Shifts of Virtual Characters. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. ACM, Santa Monica California, 1–6. https://doi.org/10.1145/2401836.2401840

Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic Gesticulator: Rhythm-Aware Co-Speech Gesture Synthesis with Hierarchical Neural Embeddings. *ACM Transactions on Graphics* 41, 6 (Dec. 2022), 1–19. https://doi.org/10.1145/3550454.3555435 arXiv:2210.01448 [cs, eess] Comment: SIGGRAPH Asia 2022 (Journal Track); Project Page: https://pku-mocca.github.io/Rhythmic-Gesticulator-Page/.

Michael Argyle and Mark Cook. 1976. *Gaze and Mutual Gaze*. Cambridge U Press, Oxford, England. xi, 210 pages.

Michael Argyle and Janet Dean. 1965. Eye-Contact, Distance and Affiliation. *Sociometry* 28 (1965), 289–304. https://doi.org/10.2307/2786027

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2002. Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication* 52, 3 (2002), 566–580. https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Birtukan Birawo and Pawel Kasprowski. 2022. Review and Evaluation of Eye Movement Event Detection Algorithms. *Sensors (Basel, Switzerland)* 22, 22 (Nov. 2022), 8810. https://doi.org/10.3390/s22228810

Sandika Biswas, Sanjana Sinha, Dipanjan Das, and Brojeshwar Bhowmick. 2021. Realistic Talking Face Animation with Speech-Induced Head Motion. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*. ACM, Jodhpur India, 1–9. https://doi.org/10.1145/3490035.3490305

Giuseppe Boccignone, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, and Raffaella Lanzarotti. 2020. On Gaze Deployment to Audio-Visual Cues of Social Interactions. *IEEE Access* 8 (2020), 161630–161654. https://doi.org/10.1109/ACCESS.2020.3021211

Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video Rewrite: Driving Visual Speech with Audio. In *Proc. SIGGRAPH*.

Julie N. Buchan, Martin Paré, and Kevin G. Munhall. 2007. Spatial Statistics of Gaze Fixations during Dynamic Face Processing. *Social Neuroscience* 2, 1 (March 2007), 1–13. https://doi.org/10.1080/17470910601043644

Ryan Canales, Eakta Jain, and Sophie Jörg. 2023. Real-Time Conversational Gaze Synthesis for Avatars. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games* (<conf-loc>, <city>Rennes</city>, <country>France</country>, </conf-loc>) *(MIG '23)*. Association for Computing Machinery, New York, NY, USA, Article 17, 7 pages. https://doi.org/10.1145/3623264.3624446

Justine Cassell, Yukiko I. Nakano, Timothy W. Bickmore, Candace L. Sidner, and Charles Rich. 2001. Non-Verbal Cues for Discourse Structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (Toulouse, France) *(ACL '01)*. Association for Computational Linguistics, USA, 114–123. https://doi.org/10.3115/1073012.1073028

Justine Cassell, Obed E Torres, and Scott Prevost. 1999. Turn taking versus discourse structure. *Machine conversations* (1999), 143–153.

Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. 2007. Predicting Human Gaze Using Low-Level Saliency Combined with Face Detection. In *Adv Neural Inf Process Syst*, Vol. 20.

Byungkuk Choi, Haekwang Eom, Benjamin Mouscadet, Stephen Cullingford, Kurt Ma, Stefanie Gassel, Suzi Kim, Andrew Moffat, Millicent Maier, Marco Revelant, Joe Letteri, and Karan Singh. 2022. Animatomy: An Animator-Centric, Anatomically Inspired System for 3D Facial Modeling, Animation and Transfer. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) *(SA '22)*. Association for Computing Machinery, New York, NY, USA, Article 16, 9 pages. https://doi.org/10.1145/3550469.3555398

Susana T.L. Chung, Girish Kumar, Roger W. Li, and Dennis M. Levi. 2015. Characteristics of Fixational Eye Movements in Amblyopia: Limitations on Fixation Stability and Acuity? 114 (2015), 87–99. https://doi.org/10.1016/j.visres.2015.01.016

Cagla Cig, Zerrin Kasap, Arjan Egges, and Nadia Magnenat-Thalmann. 2010. Realistic Emotional Gaze and Head Behavior Generation Based on Arousal and Dominance Factors. In *Motion in Games (Lecture Notes in Computer Science)*, Ronan Boulic, Yiorgos Chrysanthou, and Taku Komura (Eds.). Springer, Berlin, Heidelberg, 278–289. https://doi.org/10.1007/978-3-642-16958-8_26

J. Dawson. 2022. *Visual Attention during Conversation: An Investigation Using Real-World Stimuli.* Ph. D. Dissertation. University of Essex.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

Gwyneth Doherty-Sneddon, Deborah M Riby, and Lisa Whittle. 2012. Gaze Aversion as a Cognitive Load Management Strategy in Autism Spectrum Disorder and Williams Syndrome. *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 53, 4 (April 2012), 420–430. https://doi.org/10.1111/j.1469-7610.2011.02481.x

Andrew Duchowski, Sophie Jörg, Aubrey Lawson, Takumi Bolte, Lech Świrski, and Krzysztof Krejtz. 2015. Eye Movement Synthesis with 1/f Pink Noise. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games (MIG '15)*. Association for Computing Machinery, New York, NY, USA, 47–56. https://doi.org/10.1145/2822013.2822014

Andrew T. Duchowski, Sophie Jörg, Jaret Screws, Nina A. Gehrer, Michael Schönenberg, and Krzysztof Krejtz. 2019. Guiding Gaze: Expressive Models of Reading and Face Scanning. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications (ETRA '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3314111.3319848

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics* 35, 4 (July 2016), 1–11. https://doi.org/10.1145/2897824.2925984

Pif Edwards, Chris Landreth, Mateusz Popławski, Robert Malinowski, Sarah Watling, Eugene Fiume, and Karan Singh. 2020a. JALI-Driven Expressive Facial Animation and Multilingual Speech in Cyberpunk 2077. In *ACM SIGGRAPH 2020 Talks* (Virtual Event, USA) *(SIGGRAPH '20)*. Association for Computing Machinery, New York, NY, USA, Article 60, 2 pages. https://doi.org/10.1145/3388767.3407339

Pif Edwards, Chris Landreth, Mateusz Popławski, Robert Malinowski, Sarah Watling, Eugene Fiume, and Karan Singh. 2020b. JALI-Driven Expressive Facial Animation and Multilingual Speech in Cyberpunk 2077. In *ACM SIGGRAPH 2020 Talks* (Virtual Event, USA) *(SIGGRAPH '20)*. Association for Computing Machinery, New York, NY, USA, Article 60, 2 pages. https://doi.org/10.1145/3388767.3407339

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. 2018. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. *ACM Transactions on Graphics* 37, 4 (Aug. 2018), 1–11. https://doi.org/10.1145/3197517.3201357 arXiv:1804.03619 [cs, eess]

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. Joint Audio-Text Model for Expressive Speech-Driven 3D Facial Animation. *Proc. ACM Comput. Graph. Interact. Tech.* 5, 1, Article 16 (may 2022), 15 pages. https://doi.org/10.1145/3522615

Ylva Ferstl. 2023. Generating Emotionally Expressive Look-At Animation. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games* (, Rennes, France,) *(MIG '23)*. Association for Computing Machinery, New York, NY, USA, Article 15, 6 pages. https://doi.org/10.1145/3623264.3624438

Tom Foulsham, Joey Cheng, Jessica Tracy, Joseph Henrich, and Alan Kingstone. 2010. Gaze Allocation in a Dynamic Situation: Effects of Social Status and Speaking. *Cognition* 117 (Oct. 2010), 319–31. https://doi.org/10.1016/j.cognition.2010.09.003

Aisha Frampton-Clerk and Oyewole Oyekoya. 2022. Investigating the Perceived Realism of the Other User's Look-Alike Avatars. In *28th ACM Symposium on Virtual Reality Software and Technology*. ACM, Tsukuba Japan, 1–5. https://doi.org/10.1145/3562939.3565636

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *Computer Graphics Forum* 42, 1 (2023), 206–216. https://doi.org/10.1111/cgf.14734 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14734

Toni Giorgino. 2009. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software* 31, 7 (2009), 1–24. https://doi.org/10.18637/jss.v031.i07

A. M. Glenberg, J. L. Schroeder, and D. A. Robertson. 1998. Averting the Gaze Disengages the Environment and Facilitates Remembering. *Memory & Cognition* 26, 4 (July 1998), 651–658. https://doi.org/10.3758/bf03211385

Charles Goodwin. 1980. Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning. *Sociological Inquiry* 50, 3-4 (July 1980), 272–302. https://doi.org/10.1111/j.1475-682X.1980.tb00023.x

Ific Goudé, Alexandre Bruckert, Anne-Hélène Olivier, Julien Pettré, Rémi Cozot, Kadi Bouatouch, Marc Christie, and Ludovic Hoyet. 2023. Real-Time Multi-map Saliency-driven Gaze Behavior for Non-conversational Characters. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–13. https://doi.org/10.1109/TVCG.2023.3244679

Jennifer X. Haensel, Tim J. Smith, and Atsushi Senju. 2022. Cultural Differences in Mutual Gaze during Face-to-Face Interactions: A Dual Head-Mounted Eye-Tracking Study. *Visual Cognition* 30 (2022), 100–115. https://doi.org/10.1080/13506285.2021.1928354

D. Y. P. Henriques, W. P. Medendorp, A. Z. Khan, and J. D. Crawford. 2002. Visuomotor Transformations for Eye-Hand Coordination. *Progress in Brain Research* 140 (2002), 329–340. https://doi.org/10.1016/S0079-6123(02)40060-X

Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and Listening with the Eyes: Gaze Signaling during Dyadic Interactions. *PloS One* 10, 8 (2015), e0136905. https://doi.org/10.1371/journal.pone.0136905

Laurent Itti. 2006. Quantitative Modelling of Perceptual Salience at Human Eye Position. *Visual Cognition* 14, 4-8 (Aug. 2006), 959–984. https://doi.org/10.1080/13506280500195672

Laurent Itti, Nitin Dhavale, and Frederic Pighin. 2004. Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention. In *Optical Science and Technology, SPIE's 48th Annual Meeting*, Bruno Bosacchi, David B. Fogel, and James C. Bezdek (Eds.). San Diego, California, USA, 64. https://doi.org/10.1117/12.512618

Dušan Jan, David Herrera, Bilyana Martinovski, David Novick, and David Traum. 2007. A Computational Model of Culture-Specific Conversational Behavior. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Catherine Pelachaud, Jean-Claude Martin, Elisabeth André, Gérard Chollet, Kostas Karpouzis, and Danielle Pelé (Eds.). Springer, Berlin, Heidelberg, 45–56. https://doi.org/10.1007/978-3-540-74997-4_5

Aobo Jin, Qixin Deng, Yuting Zhang, and Zhigang Deng. 2019. A Deep Learning-Based Model for Head and Eye Motion Generation in Three-party Conversations. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2, 2 (July 2019), 1–19. https://doi.org/10.1145/3340250

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics* 36, 4 (July 2017), 1–12. https://doi.org/10.1145/3072959.3073658

Adam Kendon. 1967. Some Functions of Gaze-Direction in Social Interaction. *Acta Psychologica* 26 (Jan. 1967), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

Mohamed Amine Kerkouri and Aladine Chetouani. 2021. A Simple and Efficient Deep Scanpath Prediction. https://doi.org/10.48550/arXiv.2112.04610 arXiv:2112.04610 [cs] Comment: Electronic Imaging Symposium 2022 (EI 2022).

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1412.6980v9

Alex Klein, Zerrin Yumak, Arjen Beij, and A. Frank van der Stappen. 2019. Data-driven Gaze Animation using Recurrent Neural Networks. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games* (Newcastle upon Tyne, United Kingdom) *(MIG '19)*. Association for Computing Machinery, New York, NY, USA, Article 4, 11 pages. https://doi.org/10.1145/3359566.3360054

Krzysztof Krejtz, Andrew Duchowski, Heng Zhou, Sophie Jörg, and Anna Niedzielska. 2018. Perceptual Evaluation of Synthetic Gaze Jitter. 29, 6 (2018), e1745. https://doi.org/10.1002/cav.1745

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A Framework for Semantically-Aware Speech-Driven Gesture Generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 242–250. https://doi.org/10.1145/3382507.3418815 arXiv:2001.09326 [cs, eess] Comment: ICMI 2020 Best Paper Award. Code is available. 9 pages, 6 figures.

Binh H. Le, Xiaohan Ma, and Zhigang Deng. 2012. Live Speech Driven Head-and-Eye Motion Generators. *IEEE Transactions on Visualization and Computer Graphics* 18, 11 (Nov. 2012), 1902–1914. https://doi.org/10.1109/TVCG.2012.74

R. John Leigh and David S. Zee. 2006. The Neurology of Eye Movements.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. *CoRR* abs/1906.08172 (2019). arXiv:1906.08172 http://arxiv.org/abs/1906.08172

Laina G. Lusk and Aaron D. Mitchel. 2016. Differential Gaze Patterns on Eyes and Mouth During Audiovisual Speech Segmentation. *Frontiers in Psychology* 7 (2016).

Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué. 2009. Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos. *International Journal of Computer Vision* 82, 3 (May 2009), 231–243. https://doi.org/10.1007/s11263-009-0215-3

Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '13)*. Association for Computing Machinery, New York, NY, USA, 25–35. https://doi.org/10.1145/2485895.2485900

Anjanie McCarthy, Kang Lee, Shoji Itakura, and Darwin W. Muir. 2008. Gaze Display When Thinking Depends on Culture and Context. *Journal of Cross-Cultural Psychology* 39 (2008), 716–729. https://doi.org/10.1177/0022022108323807

Craig H. Meyer, Adrian G. Lasker, and David A. Robinson. 1985. The upper limit of human smooth pursuit velocity. *Vision Research* 25, 4 (Jan. 1985), 561–563. https://doi.org/10.1016/0042-6989(85)90160-9

Louis-Philippe Morency, C. Mario Christoudias, and Trevor Darrell. 2006. Recognizing Gaze Aversion Gestures in Embodied Conversational Discourse. In *Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI '06)*. Association for Computing Machinery, New York, NY, USA, 287–294. https://doi.org/10.1145/1180995.1181051

Atsushi Nakazawa, Yu Mitsuzumi, Yuki Watanabe, Ryo Kurazume, Sakiko Yoshikawa, and Miwako Honda. 2020. First-Person Video Analysis for Evaluating Skill Level in the Humanitude Tender-Care Technique. *Journal of Intelligent & Robotic Systems* 98, 1 (April 2020), 103–118. https://doi.org/10.1007/s10846-019-01052-8

Aline Normoyle, Jeremy B. Badler, Teresa Fan, Norman I. Badler, Vinicius J. Cassol, and Soraia R. Musse. 2013. Evaluating Perceived Trust from Procedurally Animated Gaze. In *Proceedings of Motion on Games* (Dublin 2, Ireland) *(MIG '13)*. Association for Computing Machinery, New York, NY, USA, 141–148. https://doi.org/10.1145/2522628.2522630

NVIDIA. 2021. Nemo Speaker Diarization. https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/asr/speaker_diarization/intro.html

Jason Osipa. 2010. *Stop Staring: Facial Modeling and Animation Done Right* (3rd ed.). SYBEX Inc.

Matthew K.X.J. Pan, Sungjoon Choi, James Kennedy, Kyna McIntosh, Daniel Campos Zamora, Gunter Niemeyer, Joohyung Kim, Alexis Wieland, and David Christensen. 2020. Realistic and Interactive Robot Gaze. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Las Vegas, NV, USA, 11072–11078. https://doi.org/10.1109/IROS45743.2020.9341297

Yifang Pan, Chris Landreth, Eugene Fiume, and Karan Singh. 2022. VOCAL: Vowel and Consonant Layering for Expressive Animator-Centric Singing Animation. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) *(SA '22)*. Association for Computing Machinery, New York, NY, USA, Article 18, 9 pages. https://doi.org/10.1145/3550469.3555408

Frederick I. Parke. 1998. *Computer Gernerated Animation of Faces.* Association for Computing Machinery, New York, NY, USA, 241–247. https://doi.org/10.1145/280811.281000

Tomislav Pejsa, Daniel Rakita, Bilge Mutlu, and Michael Gleicher. 2016. Authoring directed gaze for full-body motion capture. *ACM Transactions on Graphics* 35, 6 (Dec. 2016), 161:1–161:11. https://doi.org/10.1145/2980179.2982444

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022a. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/arXiv.2212.04356 arXiv:2212.04356 [cs, eess]

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022b. Robust Speech Recognition via Large-Scale Weak Supervision. http://arxiv.org/abs/2212.04356 arXiv:2212.04356 [cs, eess].

Raoudha Rahmeni, Anis Ben Aicha, and Yassine Ben Ayed. 2020. Acoustic features exploration and examination for voice spoofing counter measures with boosting machine learning techniques. *Procedia Computer Science* 176 (2020), 1073–1082. https://doi.org/10.1016/j.procs.2020.09.103

Alexander Richard, Michael Zollhoefer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. 2021. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. *arXiv:2104.08223 [cs]* (April 2021). http://arxiv.org/abs/2104.08223 arXiv: 2104.08223.

Federico Rossano. 2012. Gaze in Conversation. In *The Handbook of Conversation Analysis* (first ed.), Jack Sidnell and Tanya Stivers (Eds.). Wiley, 308–329. https://doi.org/10.1002/9781118325001.ch15

K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. 2015. A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum* 34, 6 (2015), 299–326. https://doi.org/10.1111/cgf.12603

Peiteng Shi, Markus Billeter, and Elmar Eisemann. 2020. SalientGaze: Saliency-based Gaze Correction in Virtual Reality. *Computers & Graphics* 91 (Oct. 2020), 83–94. https://doi.org/10.1016/j.cag.2020.06.007

Sinan Sonlu, Uğur Güdükbay, and Funda Durupinar. 2021. A Conversational Agent Framework with Multi-Modal Personality Expression. *ACM Trans. Graph.* 40, 1, Article 7 (jan 2021), 16 pages. https://doi.org/10.1145/3439795

Matthew Stone, Doug DeCarlo, Insuk Oh, Christian Rodriguez, Adrian Stere, Alyssa Lees, and Chris Bregler. 2004. Speaking with Hands: Creating Animated Conversational Characters from Recordings of Human Performance. In *ACM SIGGRAPH 2004 Papers* (Los Angeles, California) *(SIGGRAPH '04)*. Association for Computing Machinery, New York, NY, USA, 506–513. https://doi.org/10.1145/1186562.1015753

Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2013. Appearance-Based Gaze Estimation Using Visual Saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 2 (Feb. 2013), 329–341. https://doi.org/10.1109/TPAMI.2012.101

Justus Thies, Mohamed A. Elgharib, Ayush Tewari, C. Theobalt, and M. Nießner. 2020. Neural Voice Puppetry: Audio-driven Facial Reenactment. In *ECCV*. https://doi.org/10.1007/978-3-030-58517-4_42

J. van der Steen. 2009. Vestibulo-Ocular Reflex (VOR). In *Encyclopedia of Neuroscience*, Marc D. Binder, Nobutaka Hirokawa, and Uwe Windhorst (Eds.). Springer, Berlin, Heidelberg, 4224–4228. https://doi.org/10.1007/978-3-540-29678-2_6310

Jason Vandeventer, Andrew J. Aubrey, Paul L. Rosin, and David Marshall. 2015. 4D Cardiff Conversation Database (4D CCDb): a 4D database of natural, dyadic conversations. In *Proc. Auditory-Visual Speech Processing*. 157–162.

Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. 2021. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. https://doi.org/10.48550/arXiv.2107.09293 arXiv:2107.09293 [cs]

Nigel G. Ward, Chelsey N. Jurado, Ricardo A. Garcia, and Florencia A. Ramos. 2016. On the Possibility of Predicting Gaze Aversion to Improve Video-Chat Efficiency. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. Association for Computing Machinery, New York, NY, USA, 267–270. https://doi.org/10.1145/2857491.2857497

Justin W. Weeks, Ashley N. Howell, and Philippe R. Goldin. 2013. Gaze Avoidance in Social Anxiety Disorder. *Depression and Anxiety* 30, 8 (Aug. 2013), 749–756. https://doi.org/10.1002/da.22146

Thibaut Weise, Hao Li, Luc Van Gool, and Mark Pauly. 2009. Face/Off: live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '09*. ACM Press, New Orleans, Louisiana, 7. https://doi.org/10.1145/1599470.1599472

Sang Hoon Yeo, Martin Lesmana, Debanga R. Neog, and Dinesh K. Pai. 2012. Eyecatch: Simulating Visuomotor Coordination for Object Interception. *ACM Transactions on Graphics* 31, 4 (July 2012), 42:1–42:10. https://doi.org/10.1145/2185520.2185538

Sangbong Yoo, Seongmin Jeong, Seokyeon Kim, and Yun Jang. 2021. Saliency-Based Gaze Visualization for Eye Movement Analysis. *Sensors (Basel, Switzerland)* 21, 15 (July 2021), 5178. https://doi.org/10.3390/s21155178

Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A Large Evaluation of Data-Driven Co-Speech Gesture Generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 736–747. https://doi.org/10.1145/3536221.3558058

L.R. Young and L. Stark. 1963. Variable Feedback Experiments Testing a Sampled Data Model for Eye Tracking Movements. *IEEE Transactions on Human Factors in Electronics* HFE-4, 1 (Sept. 1963), 38–51. https://doi.org/10.1109/THFE.1963.231285 Conference Name: IEEE Transactions on Human Factors in Electronics.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In *European Conference on Computer Vision (ECCV)*.

Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. 2018. Visemenet: audio-driven animator-centric speech animation. *ACM Transactions on Graphics* 37, 4 (Aug. 2018), 1–10. https://doi.org/10.1145/3197517.3201292

Goranka Zoric, Rober Forchheimer, and Igor S. Pandzic. 2011. On Creating Multimodal Virtual Humans—Real Time Speech Driven Facial Gesturing. *Multimedia Tools and Applications* 54, 1 (Aug. 2011), 165–179. https://doi.org/10.1007/s11042-010-0526-y