Evaluating the Influence of Phoneme-Dependent Dynamic Speaker Directivity of Embodied Conversational Agents' Speech

Jonathan Ehret*
Visual Computing Institute,
RWTH Aachen University, Germany

Jonas Stienen
Institute of Technical Acoustics,
RWTH Aachen University, Germany

Chris Brozdowski
Human Technology Centre,
RWTH Aachen University, Germany

Andrea Bönsch
Visual Computing Institute,
RWTH Aachen University, Germany

Irene Mittelberg
Human Technology Centre,
RWTH Aachen University, Germany

Michael Vorländer
Institute of Technical Acoustics,
RWTH Aachen University, Germany

Torsten W. Kuhlen
Visual Computing Institute,
RWTH Aachen University, Germany

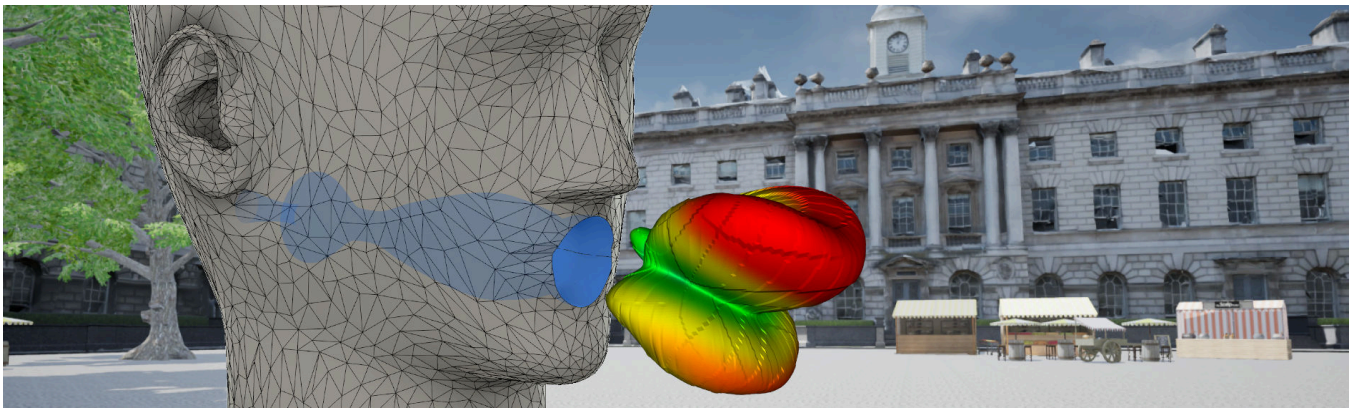


Figure 1: An exemplary speaker directivity for the vowel 'a' at 1600Hz (modulation per direction shown as distance and color (from attenuation (green) to amplification (red)) displayed in front of the used head model including vocal tract within and mouth opening (blue). In the background the used study outdoor scene can be seen.

ABSTRACT

Generating natural embodied conversational agents within virtual spaces crucially depends on speech sounds and their directionality. In this work, we simulated directional filters to not only add directionality, but also directionally adapt each phoneme. We therefore mimic reality where changing mouth shapes have an influence on the directional propagation of sound. We conducted a study ($n = 32$) evaluating naturalism ratings, preference and distinguishability of omnidirectional speech auralization compared to static and dynamic, phoneme-dependent directivities. The results indicated that participants cannot distinguish dynamic from static directivity. Furthermore, participants' preference ratings aligned with their naturalism ratings. There was no unanimity, however, with regards to which auralization is the most natural.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Virtual reality**.

KEYWORDS

embodied conversational agents, virtual acoustics, directional 3D sound, speech, phoneme-dependent directivity

ACM Reference Format:
Jonathan Ehret, Jonas Stienen, Chris Brozdowski, Andrea Bönsch, Irene Mittelberg, Michael Vorländer, and Torsten W. Kuhlen. 2020. Evaluating the Influence of Phoneme-Dependent Dynamic Speaker Directivity of Embodied Conversational Agents' Speech. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423863

1 INTRODUCTION

Embedding embodied conversational agents (ECAs) in virtual reality (VR) has become both more popular and technologically feasible,

*e-mail: ehret@vr.rwth-aachen.de

IVA '20, Oct. 19–23, 2020, Virtual Event, Scotland Uk

with increasing efforts to mimic natural human behavior. Acoustic simulations play a key role in this simulation process, including the auralization of agents' speech sounds. While spatial audio already allows users to identify the location of a speaking agent, this doesn't capture the direction the agent is facing.

In this work, we investigated the effect of adding this directional component to agents' speech on users' experience. First, we simulated static directivities, where directionality is added but does not change based on the content of the speech. Second, we simulated dynamic, phoneme-based directivities, where the directional radiation pattern changes based on the speaker's mouth shape. We conducted a study gathering preference and naturalism ratings of these *static* and *dynamic* auralizations, in contrast to *omnidirectional* audio, which emits the sound equally in all directions. Additionally, we tested whether the participants were able to reliably distinguish these auralizations. Our contribution thereby is to simulate and add dynamic directivities to ECAs' speech and evaluate them in terms of naturalism and perceptibility.

2 RELATED WORK

2.1 Spatial Acoustics

The auralization of virtual environments incorporates perceptually relevant physical effects that occur during generation, propagation and pick-up of sound. The human auditory system and the generation of binaural cues in auralization have been a research subject for decades and are investigated extensively (e.g., [7], [16]). It is well-known that directional sound must be appropriately reproduced in order to create a convincing acoustic display, which can be achieved based on binaural technology or other spatial audio formats [25]. The subjective impression of an acoustically responsive environment is modeled by simulation that applies acoustic phenomena during propagation, and integrates reflections at surface boundaries [13]. Especially in the context of virtual environments, these simulation methods are based on Geometrical Acoustics, where both sound source and receiver are considered infinitesimally small [25]. Point sources radially emit a signal from their location into space and are considered omnidirectional, since the orientation of the source is not taken into account. In order to maintain the direction-related spectral attenuation that every real-world sound source inherits, a directivity filter can be applied [19]. Directivity filtering generates perceivable differences to the user, for example, dampening the higher frequencies of speech if the speaker is looking into the opposite direction.

The directional radiation pattern of a sound source, the directivity, is commonly acquired by measurement or physics-based simulation at discrete angular directions [5]. A given directivity is *static (S)* if the spectra for all directions are not varying over time. A directivity is *dynamic (D)*, if the radiation pattern is time-variant (e.g., depends on the content of the speech). Hence, a sound object, such as a virtual agent's head, may have either a static or a dynamic directivity[1]. A neutral directivity is attributed *omnidirectional (O)* and does not modify the radiated sound.

[1]This choice of words stands in contrast to related publications (e.g., [1, 17, 18]), where dynamic directivity is attributed to a dynamic/moving sound source rotation and not a time-variant dataset.

Recent work investigated moving sound sources with static directivities and report improvements, when accounting for the varying orientation (rotational movement) of the source [1, 24].

Postma and Katz report significant differences in the room acoustics clarity and distance perception when presenting auralizations based on recordings that capture a singer's voice simultaneously at many locations and thereby naturally include directivity [17, 18]. These findings therefore encourage the use of directivities in general.

In the domain of music, Ackermann and colleagues added directivities to isolated dry recordings of instruments based on the musicians' motion, and provide evidence that listeners can reliably distinguish between static and moving auralizations [1]. It remains an open question if time-variant switching of the directivities leads to a perceivable difference compared to the averaged directivity.

During speech, the human vocal tract greatly changes. Different vowels are associated with various mouth openings, which result in variance of the directional pattern [3, 11]. Therefore, we chose to investigate the dynamics of phoneme-dependent directivities in the context of embodied conversational agents and we are interested in understanding the perceptibility of this process when simulated.

2.2 Embodied Conversational Agents

Embodied conversational agents (ECAs) are computer-controlled anthropomorphic characters, using natural language [9]. For ECAs to be believable, several components have to be simulated (e.g., lip syncing, gestures, mimics, gaze, posture, etc.), with some literature suggesting that user preference will be linked to perceived realism of ECAs, potentially including speech sound realism [29]. More work needs to be done, however, to test this claim. In the context of virtual acoustics, existing studies demonstrate that a realistic sound environment in VR can significantly improve presence, or the feeling of being there in the virtual world [20]. Presence has been bolstered, for example, by adding soundscapes or step sound [12], or a more advanced speech auralization [8]. Some research has been conducted to increase the naturalism by introducing artificial breathing during speech [6, 22], as well as phoneme-dependent lip syncing and face animations [23, 28]. Mehra and colleagues [15] suggest that sound source directivities may improve the realism of auralizations. Additionally, Wendt et al. [27] examined the influence of static directivity in ECA speech contexts with inconclusive results. The full effect of speech directivity is thus not yet understood. The present study work will try to close this gap, by comparing *dynamic, phoneme-dependent* to *static* speech directivity and *omnidirectional* auralization.

3 ACOUSTICS SIMULATION

In the context of a VR environment with an ECA as sound source, a voice signal is emitted that must be frequency-shaped according to the directional radiation pattern of the ECA's head and torso. This signal is further shaped by variations in mouth opening during speech.

3.1 Preprocessing

Directivity datasets can be either acoustically measured or simulated with physics-based approaches. The first method is commonly used for technical devices, like loudspeakers and musical instruments [26], and employs an array of microphones surrounding the

sound source recording simultaneously. A difference analysis reveals the directional pattern for the given tone or frequency, which is usually formulated as a relative change with respect to the frontal direction in an anechoic environment [21]. The measurement procedure suffers from a limited frequency range, where valid data can be acquired. Low frequencies are limited by the measurement chamber's ability to mitigate external noise, and high frequencies are limited by the spatial resolution of the measurement array. Thereby the high limit does not usually cover the full audible range. If the sound source can be described by a 3D model, the acoustic radiation can be determined by simulation, for example, with the Boundary Element Method (BEM). To simulate a phoneme-dependent directivity, a 3D head model of a human was selected and combined with different settings of a simplified tube model representing the vocal tract following Arai [3], using COMSOL Multiphysics Version 5.4 for the audible frequency range (30Hz to 16kHz in third-octave resolution). A virtual sensor array arranged in an Euler grid of 1° angular resolution acquired the transfer functions from the tube's end where the vocal chords are located. As the model is symmetrical, only one side of the head and vocal tract model was simulated. The results were post-processed by Matlab Version 2018b to normalize the filter values to the frontal direction and mirror the half-sided dataset to cover the full sphere around the sound source. Finally, the data was exported in the OpenDAFF format[2], which provides an interface to access discrete directional data stored as a lookup table.

In the end, three vowels were selected that cover the mouth opening range on the IPA vowel chart[3]. The strong similarity between directivities with the same mouth opening but different tongue position prompted us to neglect this dimension. Therefore, three representative directivity datasets for open mouth (a), half-open mouth (e) and closed mouth (i) have been simulated and all other vowels were mapped to these.

Figure 2 shows the statistical evaluation of the simulated phoneme-dependent directivities with frequency modulations for all directions and phonemes. The curves indicate moderate damping in the lower frequency range, an articulated region around 1 kHz and slow roll-off towards higher frequencies. The directivity index (DI) is a measure to reveal frontal focusing, where levels above 0 dB are stronger towards the front. Hence, both the lower and the higher frequency range of the DI curve indicate a focusing to the front and attenuation toward other directions [7]. Our mean deviations between different directions are well above the theoretically audible threshold of approximately 1 dB. In contrast, variations between vowels as indicated by the deep blue interval only show a slightly noticeable difference that exceeds 1 dB just for frequencies above 4 kHz.

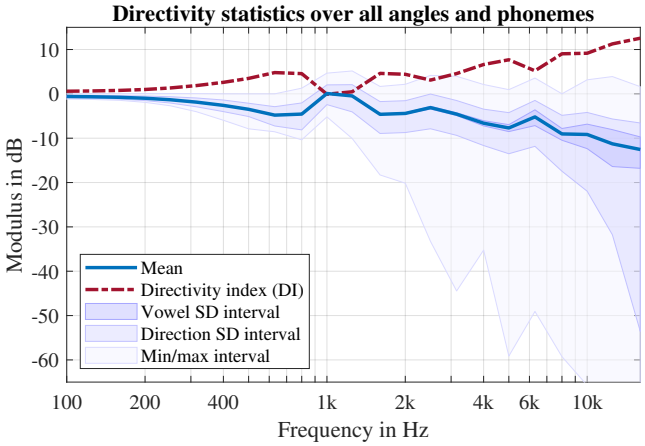


Figure 2: Simulated directivity statistics over all angles and phonemes described by Arai in [3]. SD intervals are given as mean +/- standard deviation (SD), either between the simulated vowels or between all directions.

3.2 Auralization

To render the acoustic virtual environment and to reproduce the binaural signal at the user's ears, we employed a real-time auralization framework[4]. The agent represented a dynamic, moving sound source that emitted the recorded speech signal. Directivity lookup tables were preloaded and could therefore be assigned to the sound source without delay. This way, switching between directivities had an instantaneously perceivable effect. The realization of the directivity filter bank used a transform of 256 interpolated frequency values from the third-octave resolution into the time domain, which provided an impulse response with the fingerprint of the directional attenuation. Hence, the directivity was made audible by convolving the speech signal with dynamic filter coefficients tied to the emission angle as calculated from the source's position and orientation, as well as the receiver's location. Free-field propagation simulation was applied covering properties like spherical spreading loss (amplification factor depending on relative distance), Doppler shift (resampling depending on relative movement) and the binaural filtering of the incident wave front at the receiver's ears (depending on listener's location and orientation). The binaural audio stream from the auralization pipeline was fed to a reproduction module that uses a 12-loudspeaker audio system that is mounted at the ceiling of a 5-sided CAVE. The approach is based on the multi-channel dynamic cross-talk cancellation system [14], which is able to recreate a binaural two-channel audio signal at the user's ears (also referred to as virtual headphone or transaural playback system).

If the communication between an ECA and a user is primarily face-to-face, the directivity variation in the direct sound can be expected to have little effect on the signal. Consequently in our study, different emission angles were enforced by rotation animation of the ECA, and the subjects were encouraged to move in a defined area (cf. Figure 3b). In an indoor environment, reflections off walls need to be taken into account. They can be determined according to the image source method by Allen and Berkeley [2]. If the ECA for example talks away from the user towards a wall, the damped direct sound is overlaid with the frontal speech sound reflected off the wall. To that end, we created five image sound sources for the walls and the floor. The ceiling was considered non-reflecting, due to computational limitations and it not being visible in the 5-sided CAVE. These image sound sources used the same directivity pattern as the primary source since the filters are, by design, symmetrical. This method approximates early reflections within the room, but further aspects of reverberation were not simulated due to the complexity of

[2]OpenDAFF, www.opendaff.org
[3]International Phonetic Alphabet (IPA): www.internationalphoneticassociation.org
[4]Virtual Acoustics (VA): http://www.virtualacoustics.org

(a) (b) (c)

Figure 3: The different Settings: (a) The Somerset House courtyard and (b) the museum room with a virtual agent, a museum stand and a red area the participants should move on. (c) The actor performing the full-body movements of the speech.

providing real-time update rates during dynamic directivity switching. Furthermore, the experimental environment has a reverberant characteristic that cannot be acoustically treated without interfering with other components. This reverberant characteristic is, however, comparable to the virtual room used in the study. Nevertheless, a problem arises if directivity datasets are exchanged in indoor situations because the overall energy emitted into the room changes in an non-physical way. This is most apparent when an omnidirectional directivity dataset is replaced with a human directivity dataset during auralization because the higher frequencies are dampened drastically to the back of the ECA and result in a harsh drop in that region. This effect is expected and has been intentionally included in the study to investigate the effect of energy change on naturalness ratings.

4 STUDY DESIGN

We conducted a within-subject user study in a CAVE to investigate the influence of different ECA speech Auralizations, as well as free-field *outdoor* versus acoustically reflective *indoor* Setting, on naturalism and detectability. Thereby Auralization is varied from unrealistic *omnidirectional* over adding *static* directivity to realistic phoneme-dependent *dynamic* directivity. First, an ECA took the role of a tour guide giving a 90-second speech, while allowing the participants to move freely and change Auralizations. Participants provided naturalism and preference ratings. Second, we tested whether participants were able to detect differences between the Auralizations, by pairwise testing two Auralizations in an A/B/X task.

4.1 Hypotheses

We tested the following hypotheses:

- **H1** Because *static* and *dynamic* auralizations better simulate sound propagation, participants will rate the naturalism of these conditions higher than that of an *omnidirectional* auralization.
- **H2** Because the impact of *dynamic* auralization is subtle in face-to-face settings, participants will rate the naturalism of *static* and *dynamic* directivities equally.
- **H3** Related to **H1**, participants will prefer auralizations with higher naturalism.
- **H4** Because reflections may obscure directionality cues, the differences in naturalism ratings will be stronger in the free-field *outdoor* condition compared to the *indoor* case.
- **H5** In contrast to **H2**, participants will be able to reliably distinguish *static* from *dynamic* directivity in a direct comparison.

4.2 Materials

We situated the study in a virtual version of the Somerset House in London, using a freely available scanned model from Sketchfab[5]. This model was made more lifelike, using booths and trees shown in Figure 3a. Additionally, we modeled a virtual museum room with pictures of the Somerset House as well as the model visible through the windows (cf. Figure 3b). This way the exact same speech could be given by the virtual guide in both Settings. The museum room was modeled with similar dimensions to the CAVE to match the local reverberation environment. The speech content was a 90 seconds long talk about the history and some architectural highlights of the Somerset House. An additional 30-second sequence of short words was used to feature English vowels moving from open to closed[3]. The verbal content was recorded by a native English speaker using a calibrated microphone with no frequency weighting (NTi Audio Norsonic M2230) positioned at 0.72m in front of the speaker's mouth under acoustically dry conditions. The speaker's face movements were simultaneously recorded using optical markers and a 14-camera Vicon Tracking system. Due to equipment limitations, the full body movement was recorded in a second pass wherein the speaker re-enacted the speech with co-speech gestures and head and body rotations, to showcase all parts of the directivity filters during replay (cf. Figure 3c). The movements were transferred to a virtual human model, created with Reallusion's Character Creator 3 (cf. Figure 3b), using Autodesk Motionbuilder. Unfortunately, the recorded face motion did not fit the 3D model so the lip syncing was manually re-created by an artist using Reallusion's iClone 7. Unreal Engine 4.22 was used for presentation. Furthermore, the timings of the vowels in both speeches were manually annotated and a mapping was created from all occurring vowels to the three vowels used for directivity simulation, following their position on the IPA vowel

[5]Somerset House site survey scan 2019 by Kimchi and Chips art collective: https://skfb.ly/6svNI

Figure 4: The setup of the *Detectability* task.

chart[3]. Consonants in between were auralized using the directivity of the vowel before and diphthongs were split in the middle.

4.3 Tasks

Participants engaged in two tasks: the *Comparison* and the *Detectability* task.

In the **Comparison** task, the ECA gave the above mentioned 90-second speech with associated movements in front of the participants. The speech was always identical throughout the experiment to avoid distractions. During the speech participants had the option to switch between three levels of Auralization: *omnidirectional* or featuring *static* or *dynamic* directivity. The switching was done by three dedicated buttons on the interaction device and had no perceivable delay. A sign next to the ECA displayed a letter (A, B, or C) corresponding to the auralization (see Figure 3b). Auralization-letter mapping was randomized across trials. Participants were encouraged to switch as often as they liked. Once the speech was over, a 4-item questionnaire was displayed next to the ECA, asking first "Which variant do you prefer?" Next, three 7-point Likert scales asked "How natural was Variant A/B/C" on a scale from "very unnatural"(1) to "very natural"(7). All of these questions had to be answered to continue. Alternatively, participants could answer a subset of the questions and repeat the 90-second speech. Those who repeated the trial only did so once. During the speech, participants were asked to move on a red ellipse (2.3m×1.5m) displayed underneath them, to encourage listening from different directions. The task was repeatedly performed in both Settings, *outdoor* and *indoor*, counterbalanced for order of presentation across participants.

During the **Detectability** task three equidistant ECAs were placed on platforms in front of the participants, with signs indicating A, X, and B (cf. Figure 4). If the participant clicked on a sign using a pointing ray and dedicated button, the respective agent started to speak and slowly rotate. Any other currently speaking agent was stopped. The speech content used here was a series of short words with all English vowels. Participants were told that A and B are always different and that X always matches either A or B. The participants were allowed to listen to each agent as much as they liked until they felt able to decide whether X sounded the same as A or as B. This answer was given via button press, only after each model had been played at least once.

4.4 Procedure

Participants first gave informed consent, provided demographic information and read task descriptions. Next, they entered the CAVE and performed a practice trial of the *Comparison* task, which included virtual interface instructions as well as the standard post-trial questionnaire. The practice trial compared the omnidirectional auralization to two very artificial sounding high- and low-pass filtered auralizations, with the expectation that participant would prefer the omnidirectional one. Participants who preferred the artificial auralizations were excluded from analyses. Experimental *Comparison* trials were blocked for *outdoor* versus *indoor* Setting, randomized for order of presentation across participants. For each of three experimental trials within each Setting, each Auralization was randomly paired with A, B or C. After six *Comparison* trials, participants were asked to answer the 5-item Social Presence Survey [4] and leave the CAVE for a 5-minute break. Next, participants returned to the CAVE for nine trials of the A/B/X *Detectability* task, which presented each possible combination of Auralization pairs three times in randomized order. All participants performed the *Comparison* task prior to the *Detectability* task to avoid direct comparisons influencing our measures of naturalism and preference. Finally, participants filled out a post-study questionnaire, addressing the ease of the *Detectability* task, as well as the intensity and realism of the audio, on 7-point Likert scales.

4.5 Measures

In addition to the in-VR decisions and questionnaire responses, we recorded the position and orientation of the participant and the currently speaking ECA. From these, we summed up the distance participants' heads moved between two frames during the *Comparison* task, as well as the change in emission angle. The emission angle is the angle under which the speech sound source was heard, relative to the ECA's head, and can be computed as:

$$\measuredangle_{\text{emission}} = \measuredangle\left(\ \mathbf{p}_{\text{participant}} - \mathbf{p}_{\text{agent}},\ \ \vec{d}_{\text{agent}}\ \right)$$

with the position $\mathbf{p}$ and forward direction $\vec{d}$ of the participant's and agent's head. We compute the maximum emission angle encountered during each of the Auralizations in every trial, as well as the sum of emission angle changes between all adjacent frames.

4.6 Equipment

The study was conducted in a five-sided CAVE (four walls and a floor) with a size of 5.25m × 5.25m × 3.30m ($w \times d \times h$). The participants wore tracked active stereo glasses and interacted with an ART Flystick 2. The open ceiling of the CAVE was besides the tracking system equipped with an audio system generating binaural audio using cross-talk cancellation (cf Section 3.2). Furthermore, two small surveillance cameras and microphones were mounted in the ceiling, and used by the experimenter to monitor progress.

4.7 Participants

32 participants (9 female) were primarily recruited via university mailing lists. Three participants were excluded from all analyses

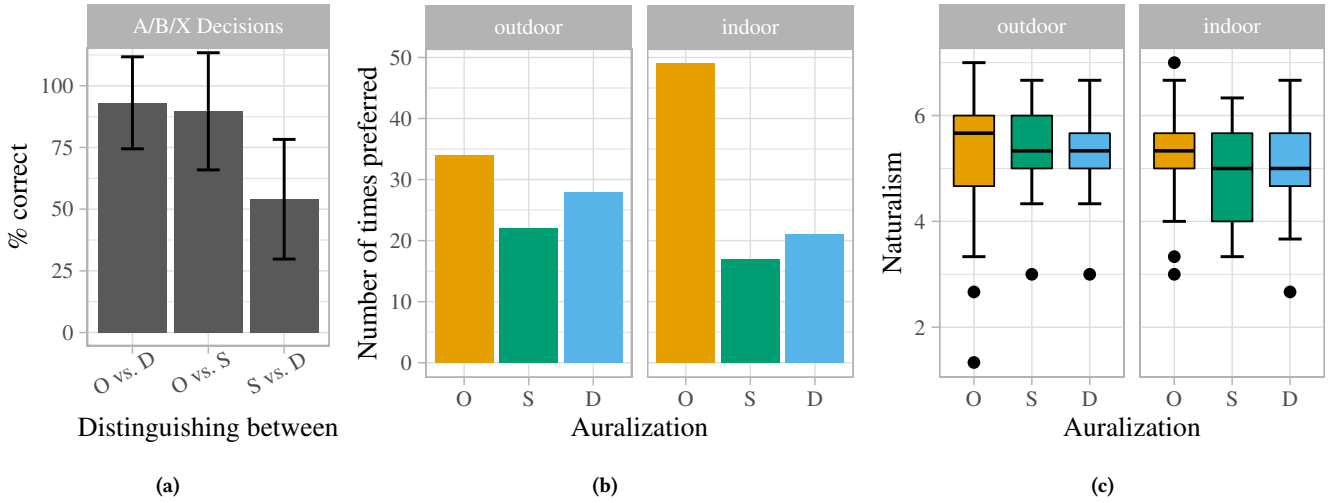


Figure 5: (a) The percentage of correct answers when trying to differentiate between the different Auralizations *omnidirectional* (O), *static* (S) and *dynamic* (D) in an A/B/X comparison. Error bars show the standard deviation. (b) The number of preferences per Auralization split by the two Settings: the courtyard scene (*outdoor*) and the museum scene (*indoor*). (c) Box plot of naturalism ratings per Setting and Auralization averaged over 3 trials based on a 7-point Likert scale from "very unnatural"(1) to "very natural"(7). Boxes indicate quartiles, with whiskers at full range.

for failure to select the correct response during the practice trial. Furthermore, answers to three *Comparison* task trials were removed where the participant failed to listen to all three auralizations. The remaining 29 participants (8 female) had a mean age of 28.24 (standard deviation (SD) = 9.85), and all reported normal hearing, normal or corrected vision and some English skills (2 "basic", 27 "fluent").

5 STUDY RESULTS

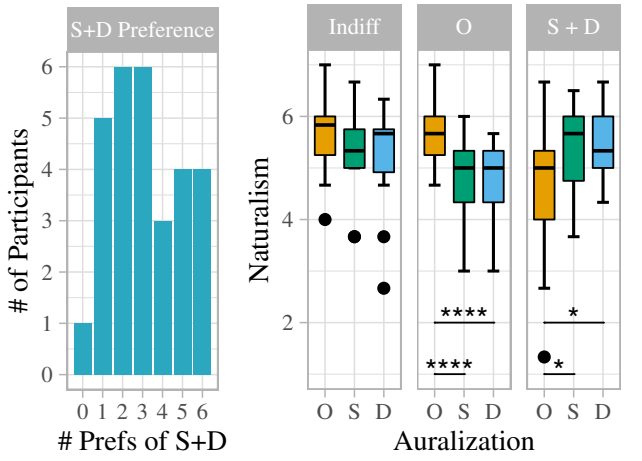
The participants rated the sound's realism on a 7-point Likert Scale from "not at all"(1) to "a lot"(7) (M = 5.00, SD = 1.10) and the sound intensity on a scale from "too silent"(1) to "too loud"(7) (M = 4.07, SD = .60) overall appropriate. Furthermore the ease of *Detectability* phase was rated on a 7-point Likert Scale from "not at all"(1) to "a lot"(7) (M = 3.65, SD = 1.78) as reasonable. Social Presence was rated with a neutral .1 (SD = 4.08) of the possible scale from -15 to 15.

The results of the *Detectability* task can be seen in Figure 5a. A planned chi-square test of independence was performed to examine the relation between the different A/B/X-comparison-pairs and the ability to detect whether X was the same Auralization as A or B. There was a significant difference between response accuracy across comparison pairs, $\chi^2(2, N = 261) = 48.97$, $p < .0001$. Furthermore two-sided binomial tests showed that *omnidirectional (O)* can be distinguished from both *static (S)* and *dynamic (D)* speaker directivity significantly better than by chance, ps $< .0001$. The A/B/X decisions between *S* and *D* speaker directivity, on the other hand, were not significantly different from chance (.5), $p = .52$.

Figure 5b highlights the frequency of preference selection in the *Comparison* task. A planned chi-square test did not reveal a main effect of the Setting: *outdoor* vs. *indoor*, $\chi^2(2, N = 171) = 4.30$, $p = .116$. The relation between preference and naturalism ratings of the different Auralizations was further analysed using a planned two-way repeated measures ANOVA. There was a statistically significant interaction between preference and Auralization on naturalism ratings per trial, $F(4, 504) = 39.6$, $p < .0001$. Therefore, the naturalism ratings per Auralizations were analysed for each preference rating. P-values were adjusted using the Bonferroni multiple testing correction method. The effect of preference was significant for all three preference outcomes (ps $< .0001$). Pairwise comparisons, using paired t-test, show that the naturalism rating of the preferred Auralization is always significantly higher than the ones of the other two Auralizations (ps $< .0001$). Furthermore the differences between the non-preferred Auralizations were non-significant(*S* vs. *D* when choosing *O* ($p = 1.0$), *O* vs. *D* when choosing *S* ($p = .13$) and *O* vs. *S* when choosing *D* ($p = .09$)) .

Figure 5c shows the mean values of the naturalism ratings, averaged over the 3 trials per Setting. A planned 2x3 (Settings by Auralizations) repeated measures ANOVA revealed no significant effects (Fs $< .932$, ps $> .40$) and only a marginal trend for Setting ($F(1, 168) = 3.231$, $p = .074$), with slightly higher naturalism ratings for *outdoor*.

Looking at the preferences of single participants (cf. Figure 6a), we saw that the number of times single participants preferred either Auralization seems to be bi-modal. We combined *S* and *D* here following the results of the *Detectability* task. This way we found that there is a group of participants strongly preferring *S + D* while some other participants were consistently in favor for *O* with a noticeable gap in between. Due to this observation we split the population into three groups: those preferring *O* (N=12) or *S + D* (N=11) more often and individuals preffering these two auralization (groups) equally often (N=6), labeled as 'Indiff'.



(a) (b)

Figure 6: (a) How many participant preferred *S+D* how often. Following the results of the *Detectability* task, *S* and *D* are combined. (b) Mean naturalism ratings per preference type (where *O* and *S+D* preferred this Auralization more than 3 times and *Indiff* both exactly 3 times) and Auralization averaged over 3 trials based on a 7-point Likert scale from "very unnatural"(1) to "very natural"(7).

A post-hoc 3x3 repeated measures ANOVA showed a significant interaction between this preference grouping (further called Preference Type) and Auralization, $F(4, 165) = 8.58$, $p < .0001$. Therefore, the effect of the Auralization was analyzed for each Preference Type. P-values were adjusted using the Bonferroni multiple testing correction method. The effect of Auralization was significant for O-preferred ($p < .001$) and S+D-preferred ($p = .009$) but not for the indifferent group ($p = 1$). Pairwise comparisons, using paired t-test, show that the naturalism ratings were significantly higher for the preferred Auralization in these two groups (O-preferred: ps $< .001$; S+D-preferred: *O* vs. *S* $p = .02$ and *O* vs. *D* $p = .01$), while no other differences were significant (ps $> .39$).

In our next post-hoc analysis, we examined participant movement during the *Comparison* task while the agent was speaking. Changes in auralization become most noticeable when the participant experiences a variety of emission angles. Participants moved an average of 123m (SD = 86m) with an accumulated change of emission angle of $13,814°$ (SD= $3,657°$). The mean maximum emission angle per Auralization and trial was $104°$ (SD = $37°$). However, there was no correlation evident in the data between this maximum emission angle and individual naturalism ratings per Auralization and trial ($p = .13$). There was also no significant correlation between the accumulated movements and naturalism ratings grouped by Auralization (ps $> .22$), with only a marginal negative correlation between accumulated emission angle change and the naturalism rating of *O*, $r(56) = -.24$, $p = .067$. Furthermore, there was also no significant correlation regarding movement and the quantity of preferences for either Auralization (ps $> .19$).

6 DISCUSSION

The results of the *Detectability* task show that participants were able to differentiate ECAs having an *omnidirectional* auralization from those using a directional one. However, participants were unable to reliably distinguish between *static* and *dynamic* directivity, even when given control over the rotation of the agents. This pattern contradicts hypothesis **H5**. However given the subtle differences in the simulated directivity data of the different vowels (cf. Figure 2), participants inability to distinguish both directional auralizations is unsurprising.

Furthermore, we were able to support **H3**. Participants rated the naturalism of their preferred auralization significantly higher than that of the other two. This aligns with proposals that listeners prefer naturally sounding ECAs [29], and affirms the need to optimize for higher naturalism.

In general, we did not observe higher naturalism ratings for the two directional auralization methods, leading us to reject **H1**. This can potentially be attributed to the fact that the low-pass filters applied in the directional auralizations in cases of lateral or even backward emission angles may detract from their perceived naturalism. While directional auralizations may be closer to reality, users seem to prefer that speech is unchanged. This finding allows for speculation about the role of low-pass filtering on speech intelligibility [10]. We also found, however, that several participants consistently select either strongly in favor of the directional auralizations or the omnidirectional auralization, while others remained indifferent (cf. Figure 6a). We therefore performed a post-hoc split based on this preference, and found significant differences in the naturalism ratings in favor of participants' preferred auralization (cf. Figure 6b).

Even under this split, there was no significant difference between *static* and *dynamic* auralizations. One possible explanation is that participants were not listening enough to non-frontal directions, as the virtual environment implied face-to-face communication. The measured maximum emission angles, however, exhibit sufficiently large movements as to elicit directional effects (mean above $90°$). Furthermore, a post-hoc analysis did not show any significant relationship between participant movement and naturalism or preference ratings. A marginal trend revealed that participants who moved more (i.e., those who had a larger summed emission angle, indicating effort to detect differences), rated the naturalism of *omnidirectional* slightly lower. We take this result to suggest that our study circumvented the possible impact of insufficient movement, which would have otherwise hidden auralization differences at non-frontal emission angles. Therefore, we cautiously support **H2**.

Given these results, it may not be necessary for ECAs to implement dynamic, phoneme-dependent directivity for a comparable face-to-face interaction. In our case, the added effort over a static directivity seems unrecognized by the listeners. This would make auralizing ECAs' voices easier, eliminating the extra link between the animation system and the acoustical simulation to switch directivity filters based on the currently uttered phoneme.

We were unable to confirm **H4**, pertaining to the strength of naturalism differences across free-field *outdoor* versus reflective *indoor* Setting. We expected that the different directivity spectra per propagation path of the additional indoor reflections would entail

perceivable deterioration of the directional effect. Furthermore, the energy mismatch caused by the missing damping of an *omnidirectional* source results in an unnatural room acoustic characteristic. Contrastingly, participants rated naturalism *outdoor* slightly higher where these effects did not occur. As no significant differences were found, however, we cannot make any conclusions about changes related to *indoor* reflections. The real reverberation of the inside cavity of the CAVE, where the walls and floor have acoustically problematic properties, may have interfered with the successful investigation of this effect. Future studies might be able to better control the experimental environment acoustically.

7 CONCLUSION

In the present study, we found hints that the integration of dynamic, phoneme-dependent directivities was not distinguishable from a static (averaged) speaker directivity. We therefore suggest that the additional effort to switch directivities during speech is not required, and we expect this to hold for comparable scenarios. Furthermore, we found no evidence that participants prefer directional speech sound in general. While nearly half of our participants preferred the auralization including directivities, there were also many participants with a strong preference for omnidirectional speech. The fact that those groups consistently reported their respective preference gives rise to the notion that subjective preference is more related to other factors not considered here (e.g., speech perception) than the realism of directional rendering. Also, future studies would be well served by using a more acoustically controlled environment, with a more robust reproduction (i.e., headphones), to better distinguish reflective and free-field settings. Nevertheless, we found that participants generally preferred the auralizations they rated as more natural, which affirms the need for higher naturalism in speech auralization.

ACKNOWLEDGMENTS

Kindly funded by RWTH Aachen University as Exploratory Research Space Seed Fund Project OPSF459. Special thanks goes also to Prakaiwan Vajrabhaya, Imran Muhammad, Lukas Aspöck, Benjamin Weyers, Mark Müller-Giebeler and Henry Andrew.

REFERENCES

[1] David Ackermann, Christoph Böhm, Fabian Brinkmann, and Stefan Weinzierl. 2019. The Acoustical Effect of Musicians' Movements During Musical Performances. *Acta Acust united Ac* 105, 2 (2019), 356–367. https://doi.org/10.3813/AAA.919319

[2] Jont B. Allen and David A. Berkley. 1979. Image Method for Efficiently Simulating Small-room Acoustics. *J. Acoust. Soc. Am.* 65 (1979), 943–950. https://doi.org/10.1121/1.382599

[3] Takayuki Arai. 2001. The Replication of Chiba and Kajiyama's Mechanical Models of the Human Vocal Cavity. *J. Phonetic Soc. Jpn.* 5, 2 (2001), 31–38. https://doi.org/10.24467/onseikenkyu.5.2_31

[4] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. 2001. Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments. *Presence: Teleoperators and Virtual Environments* 10, 6 (dec 2001), 583–598. https://doi.org/10.1162/105474601753272844

[5] Gottfried Behler, Martin Pollow, and Michael Vorländer. 2012. Measurements of Musical Instruments with Surrounding Spherical Arrays. In *Proc. Acoustics Nantes Conf.* 761–765. https://hal.archives-ouvertes.fr/hal-00811213/

[6] Ulysses Bernardet, Sin-Hwa Kanq, Andrew Feng, Steve Dipaola, and Ari Shapiro. 2019. Speech Breathing in Virtual Humans: An Interactive Model and Empirical Study. In *IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE).* 1–9. https://doi.org/10.1109/VHCIE.2019.8714737

[7] Jens Blauert. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT press.

[8] Christina Dicke, Viljakaisa Aaltonen, Anssi Rämö, and Miikka Vilermo. 2010. Talk to me: The Influence of Audio Quality on the Perception of Social Presence. In *Proc BCS HCI.* 309–318. https://dl.acm.org/doi/abs/10.5555/2146303.2146349

[9] Jonathan Gratch, Jeff Rickel, Elisabeth André, Justine Cassell, Eric Petajan, and Norman Badler. 2002. Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intell. Sys.* 17, 4 (2002), 54–63. https://doi.org/10.1109/MIS.2002.1024753

[10] Peter Jax and Peter Vary. 2006. Bandwidth Extension of Speech Signals: A Catalyst for the Introduction of Wideband Speech Coding? *IEEE Commun. Mag.* 44, 5 (may 2006), 106–111. https://doi.org/10.1109/MCOM.2006.1637954

[11] Brian F. G. Katz, Fabien Prezat, and Christophe D'Alessandro. 2006. Human Voice Phoneme Directivity Pattern Measurements. In *J. Acoust. Soc. Am.*, Vol. 120. 3359–3359. https://doi.org/10.1121/1.4781486

[12] Angelika C. Kern and Wolfgang Ellermeier. 2020. Audio in VR: Effects of a Soundscape and Movement-Triggered Step Sounds on Presence. *Front. Robot. AI* 7 (feb 2020), 20. https://doi.org/10.3389/frobt.2020.00020

[13] Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. 2007. Virtual Reality System with Integrated Sound Field Simulation and Reproduction. *EURASIP J. Adv. Sig. Pr.* 1 (dec 2007), 1–19. https://doi.org/10.1155/2007/70540

[14] Bruno Masiero and Michael Vorländer. 2014. A Framework for the Calculation of Dynamic Crosstalk Cancellation Filters. *IEEE T. Audio Speech* 22, 9 (2014), 1345–1354. https://doi.org/10.1109/TASLP.2014.2329184

[15] Ravish Mehra, Lakulish Antani, Sujeong Kim, and Dinesh Manocha. 2014. Source and Listener Directivity for Interactive Wave-based Sound Propagation. *IEEE T. Vis. Comput. Gr.* 20, 4 (2014), 495–503. https://doi.org/10.1109/TVCG.2014.38

[16] Florian Pausch, Lukas Aspöck, Michael Vorländer, and Janina Fels. 2018. An Extended Binaural Real-Time Auralization System With an Interface to Research Hearing Aids for Experiments on Subjects With Hearing Loss. *Trends Hear.* 22 (2018), 1–32. https://doi.org/10.1177/2331216518800871

[17] Barteld N. J. Postma, Hugo Demontis, and Brian F. G. Katz. 2017. Subjective Evaluation of Dynamic Voice Directivity for Auralizations. *Acta Acust united Ac* 103, 2 (2017), 181–184. https://doi.org/10.3813/AAA.919045

[18] Barteld N. J. Postma and Brian F. G. Katz. 2016. Dynamic Voice Directivity in Room Acoustic Auralizations. In *Germ. Ann. Conf. Acoustics (DAGA).* 352–355.

[19] Jens H. Rindel, Felipe Otondo, and Claus L. Christensen. 2004. Sound Source Representation for Auralization. In *Int. Symp. Room Acoustics: Design and Science.*

[20] Stefania Serafin, Michele Geronazzo, Cumhur Erkut, Niels C. Nilsson, and Rolf Nordahl. 2018. Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions. *IEEE Comput. Graph.* 38, 2 (mar 2018), 31–43. https://doi.org/10.1109/MCG.2018.193142628

[21] Noam R. Shabtai, Gottfried Behler, Michael Vorländer, and Stefan Weinzierl. 2017. Generation and Analysis of an Acoustic Radiation Pattern Database for Forty-one Musical Instruments. *J. Acoust. Soc. Am.* 141, 2 (2017), 1246–1256. https://doi.org/10.1121/1.4976071

[22] Éva Székely, Gustav E. Henter, Jonas Beskow, and Joakim Gustafson. 2020. Breathing and Speech Planning In Spontaneous Speech Synthesis. In *IEEE ICASSP.* 7649–7654. https://doi.org/10.1109/ICASSP40776.2020.9054107

[23] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio G. Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A Deep Learning Approach for Generalized Speech Animation. *ACM T. Graphic.* 36, 4 (2017), 1–11. https://doi.org/10.1145/3072959.3073699

[24] Michelle C. Vigeant, Lily M. Wang, and Jens H. Rindel. 2011. Objective and Subjective Evaluations of the Multi-channel Auralization Technique as Applied to Solo Instruments. *Appl. Acoust.* 72, 6 (2011), 311–323. https://doi.org/10.1016/j.apacoust.2010.10.004

[25] M Vorländer. 2011. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality.* Springer Berlin Heidelberg.

[26] Stefan Weinzierl, Michael Vorländer, Gottfried Behler, Fabian Brinkmann, Henrik von Coler, Erik Detzner, Johannes Krämer, Alexander Lindau, Martin Pollow, Frank Schulz, and Noam R. Shabtai. 2017. A Database of Anechoic Microphone Array Measurements of Musical Instruments. https://doi.org/10.14279/depositonce-5861.2

[27] Jonathan Wendt, Benjamin Weyers, Jonas Stienen, Andrea Bönsch, Michael Vorländer, and Torsten W. Kuhlen. 2019. Influence of Directivity on the Perception of Embodied Conversational Agents' Speech. In *Proc. Int. Conf. Intell. Virtual Agents.* ACM, 130–132. https://doi.org/10.1145/3308532.3329434

[28] Yang Zhou, Zhan Xu, Chris Landreth, Avangelos Kalogerakis, Subhransu Maji, and Karan Singh. 2018. VisemeNet: Audio-Driven Animator-Centric Speech Animation. *ACM T. Graphic.* 37, 4 (2018). https://doi.org/10.1145/3197517.3201292

[29] Katja Zibrek, Sean Martin, and Rachel McDonnell. 2019. Is Photorealism Important for Perception of Expressive Virtual Humans in Virtual Reality? *ACM T. Appl. Percept.* 16, 3 (2019). https://doi.org/10.1145/3349609