Gonzalez Diaz, Raimundo; McKenzie, Thomas; Politis, Archontis; Lokki, Tapio

Near-Field Evaluation of Reproducible Speech Sources

Published in:
Journal of the Audio Engineering Society

DOI:
10.17743/jaes.2022.0022

Published: 01/07/2022

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Near-Field Evaluation of Reproducible Speech Sources

RAIMUNDO GONZALEZ,[1] THOMAS MCKENZIE,[1] ARCHONTIS POLITIS,[2] AND TAPIO LOKKI[1]

(Submitted to J. Audio Eng. Soc., 2022 January, Revised & Resubmitted in March 2022)

[1]*Department of Signal Processing & Acoustics, Aalto University, Espoo, Finland.*
[2]*Audio & Speech Processing Group, Tampere University of Technology, Tampere, Finland.*

The spatial speech reproduction capabilities of a KEMAR mouth simulator, a loudspeaker, the piston on sphere model and a circular harmonic fitting are evaluated in the near-field. The speech directivity of 24 human subjects, both male and female, is measured using a semi-circular microphone array of radius 36.5 cm in the horizontal plane. Impulse responses are captured for the two devices and filters are generated for the two numerical models to emulate their directional effect on speech reproduction. The four repeatable speech sources are evaluated through comparison to the recorded human speech both objectively, through directivity pattern and spectral magnitude differences, and subjectively, through a listening test on perceived coloration. Results show that the repeatable sources perform relatively well under the metric of directivity but irregularities in their directivity patterns introduce audible coloration for off-axis directions.

0 INTRODUCTION

A primary application of virtual and augmented reality is that of telepresence, a technology that allows two or more individuals in separate physical locations to communicate as if they were in the same space [1]. A core requirement for immersive telepresence is the plausible reproduction of human speech, not only in its intelligibility but also in the spatial properties of its sound field. For reproducing spatially plausible human speech, there are two primary methods: using an electroacoustic driver capable of emulating the sound field produced by speech, such as a mouth simulator or a loudspeaker, which would allow the speech source to be present in real life, or synthesizing the directivity of the speech numerically, and presenting the speech source over headphones.

In immersive applications which try to emulate the acoustics of peripersonal scenes, scenarios where individuals are spatially within "arms reach" of one another, the near-field of speech is of particular interest [2]. Sounds presented in the peripersonal space have been proven to strengthen an individual's engagement with audiovisual stimuli [3]. Hence, reliable and plausible methods for reproducing near-field speech could improve intimacy [4] in telepresence and virtual reality applications, but remains uninvestigated.

In this study, the near-field properties, namely directivity and speech coloration, are evaluated for two speech reproduction devices and two numerical methods. The devices evaluated are a GRAS 45BC KEMAR Head & Torso with Mouth Simulator and a Genelec 8020B Loudspeaker. The numerical methods are the analytical model of a piston on a sphere and a measurement-based model generated by fitting circular harmonics on measured directivity data. The KEMAR mouth simulator has been chosen for its detailed anthropomorphic design, a design that is intended to repeatedly reproduce speech recordings which include the near-field self-scattering effects from the head and chest of a human speaker [5]. The loudspeaker has been chosen as a more accessible option due to its widespread use in acoustic measurements. It is known for a having a smooth directivity and has a relatively low cost when compared to laboratory mouth simulators. The piston on sphere model (PoS) is a physical model commonly used for studying speech directivity [5, 6]. It includes a piston, approximating the mouth source, mounted on a rigid spherical baffle, approximating the head. Far-field and near field directivity and scattering effects are easily parameterized in the model, in terms of the diameter of the mouth/piston and head/sphere. Finally, the measurement-based model makes use of real speech directivity data acquired in the horizontal plane, averaged over many subjects, on which circular harmonics are fitted in order to synthesize directivity filters efficiently during run-time. This model is herein termed the circular harmonic (CH) decomposition model.

This paper is laid out as follows. Section 1 presents a background on repeatable speech in the near-field, provid-

ing further motivation for this study. In Section 2, near-field human speech is recorded for 24 individuals (12 male, 12 female), followed by directivity measurements of the GRAS 45BC KEMAR with mouth simulator and a Genelec 8020B loudspeaker, performed in the same measurement setup. In addition, directivities are also synthesized from the analytical PoS model and a CH decomposition process, as potential "virtual only" repeatable speech sources. Section 3 then presents an evaluation of the repeatable speech sources. An objective evaluation is first performed through comparison to an average of the recorded human speech, using directivity measurements and coloration. A perceptual evaluation is then performed through a listening test to compare between the various methods and evaluate the audibility of differences. The results of the evaluation are discussed in Section 4, with the best suited near-field repeatable speech source established, before the paper is concluded, along with further work, in Section 5.

1 BACKGROUND

The development of reproducible speech devices is intrinsically connected to the initial studies on the sound field of speech around the human head, or speech directivity [7, 8]. A considerable amount of research has been published relating to speech directivity including large directivity samples [9], directivity while singing [10, 11, 5], the directivity of low and high levels of speech [11], the directional characteristics of specific phonemes [12, 13], as well as modern methods to capture [14, 5] and spatially up-sample [15] speech directivity.

In order to aid these studies, the anthropomorphic mouth simulator and piston on sphere model was proposed [8] and improved [16, 17]. The mouth simulator was eventually standardized [18] as a repeatable measuring device for the near and far-field reproduction of speech. The mouth simulator and piston on sphere model have been used to aid the development of speech telecommunications systems [6], measure speech intelligibility [19] and simulate one's own speech in a room [20]. The mouth simulator is usually integrated into a head and torso simulator which also includes binaural recording capabilities [21], and is used extensively in the field of hearing research and in the development of mobile phones. The piston on sphere model has been used for the study of loudspeaker design [22].

A comparison was performed between the directivity of average speech and a mouth simulator [19] in the far-field, but it only studied a single male subject up to frequencies of 2 kHz. Comparisons have also been made between the Brüel and Kjaer 4128 head and torso simulator (HATS), the PoS model and specific mouth configurations of singers, revealing good directivity matching for one normal mouth configuration [5]. The KEMAR mouth simulator has been compared to the PoS model in terms of directivity up to 6.5 kHz for four frequency bands [23]. Furthermore, the KEMAR 45BC was shown as capable of reproducing speech authentically [24], while using a larger sample of subjects as well as more objective evaluation methods. However, it only investigated the frontal direction and not directivity. This paper does not focus on the performance of the mouth simulator for the purposes of speech intelligibility, as mouth simulators have already been proven to work adequately for these kinds of tasks [14].

Methods for acoustic holography [25] used for capturing, analyzing, and emulating the radiation properties of sources [26] have been applied to the directivity of speech using arrays of microphones [13, 15, 14]. By decomposing the discrete pressure surface emanating from a speaker using a set of orthonormal basis of functions, the continuous definition of speech directivity patterns can be reconstructed up to some spatial aliasing limit.

Human spatial hearing acts differently in the near-field. Though interaural time differences are relatively unchanged in the near-field, interaural level differences become much greater below 1 m due to the increase in acoustic shadowing caused by the head [27], as do spectral deviations [28]. However, it remains uninvestigated how the directivity of reproducible speech sources acts in the near-field. It may be that differences in size, geometry and acoustical radiation patterns greater accentuate the directional differences between repeatable speech sources and human speakers, meaning it is harder to achieve authentic speech reproduction in the near-field.

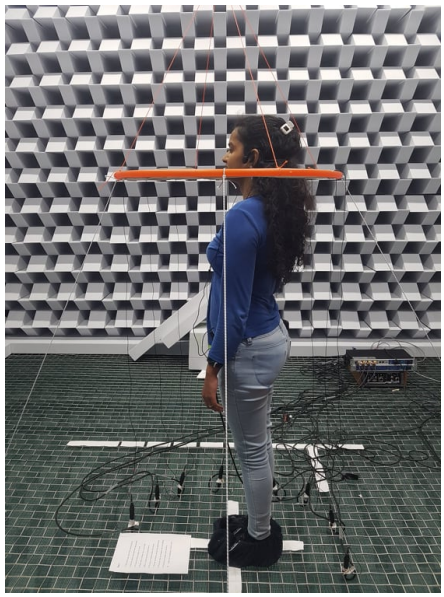
In this study, speech directivity is evaluated using monoaural signals. It was not physically possible to capture both directivity of a speech signal and its binaural recordings simultaneously, as this would require having an array of microphones and a binaural dummy-head in the same exact position at the same time. Nevertheless, the monoaural properties captured in this study should reflect the spatial perception of the studied near-field speech sources sufficiently, as in previous studies [6, 5].

2 METHODS

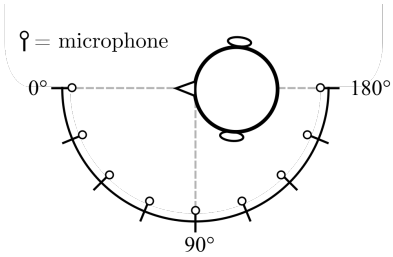
2.1 Reproducible Speech Measurements

To assess the near-field performance of the KEMAR mouth simulator and Genelec loudspeaker as repeatable speech sources, near-field speech of human speakers and near-field impulse responses (IRs) of the devices were captured for multiple directions.

The measurement setup involved a semi-cylindrical microphone array suspended inside a fully anechoic room. The structure supporting the microphone array, presented in Fig. 1a, was a 36.5 cm radius plastic ring which hung horizontally from the ceiling. The vertical height of the ring was adjustable through a pulley system and calibrated to be horizontal using a Bosch GLL 3-80 self-leveling line laser. The ring was also tied to the floor through elastic rubber cords which prevented it from moving sideways while also absorbing any potential vibrations from the floor. The dimensions of the ring were chosen because they allowed standing subjects to comfortably fit inside while also allowing to record speech at a distance significantly within the near-field.

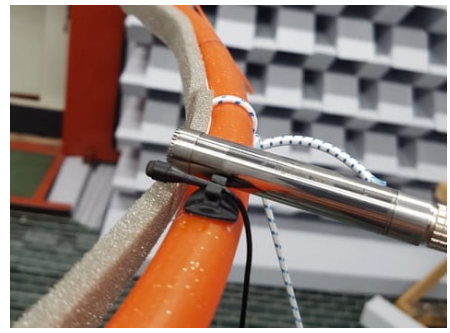


(a) Participant during speech recording

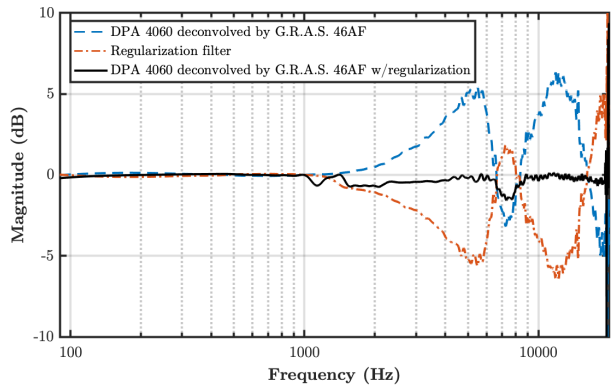


(b) DPA microphone placement

Fig. 1: Anechoic measurement setup for recording human speech using a vertically-adjustable semi-cylindrical DPA microphone array.



(a) Calibration setup



(b) The measured, inverted and resulting equalized frequency response

Fig. 2: Calibration of the DPA 4060 microphones using the GRAS 46AF measurement microphone.

Though some evidence indicates that directivity of fricative sounds can be slightly asymmetric [13], the directivity of speech was assumed to be symmetric along the forward-looking direction [5, 14], and therefore speech was captured for a semicircle in the horizontal plane. On the ring, nine DPA 4060 miniature omnidirectional microphones were arranged from a 0° direction, in 22.5° increments, until 180° as shown in Fig. 1b. The number of microphones was chosen based on the requirements for a 7^{th} order CH decomposition, as described further in Section 2.3. To minimize reflections from the ring itself, the microphones protruded inwards from the ring by 1 cm and the ring itself was covered with absorption material. The microphone array signals were interfaced into a computer at a 48 kHz sampling rate through a RME Fireface UFX+ and an RME Octomic interface synced together via a MADI connection. In the computer, the signals were recorded in the Cockos Reaper DAW and then exported to MATLAB for processing and analysis.

The frequency response of the DPA microphones was equalized to match the response of a reference G.R.A.S. 46AF 1/2" free-field measurement microphone with a flat frequency response [29]. Firstly, both microphones were aligned, as demonstrated in Fig. 2a, and an IR was measured for both microphones using a small 1 1/2" loudspeaker at the center of the ring with the swept-sine technique [30]. Next, the response of the reference microphone was deconvolved from the microphones in the array, and finally equalized using Kirkeby regularization [31] with in- and out-band regularization parameters of 10 dB and 24 dB respectively between 1.5 and 16 kHz, with no octave smoothing, as shown in Fig. 2b. The levels of all array microphones were matched to within $\pm$ 0.5 dB.

For the repeatable speech source measurements, the devices were centered in the microphone array according to their acoustic center, which for the subjects and KEMAR was defined to be the mouth [5], and for the loudspeaker its frontal baffle. This approach minimizes spatial aliasing effects in the CH decomposition approach described in Section 2.3 due to reduced time-of-arrival differences between microphone signals [26].

2.1.1 Anechoic Speech Recording

The speech of 24 human subjects (12 male and 12 female) was recorded using the microphone ring in an anechoic chamber at Aalto University, Espoo, Finland. Of all recorded subjects, only 4 were native English speakers. As shown in Fig. 1a, subjects stood inside the ring facing the 0° microphone, with their mouths laser-aligned to the center of the array. Subjects stood up to avoid potential reflections from their laps. The height of the ring was ad-

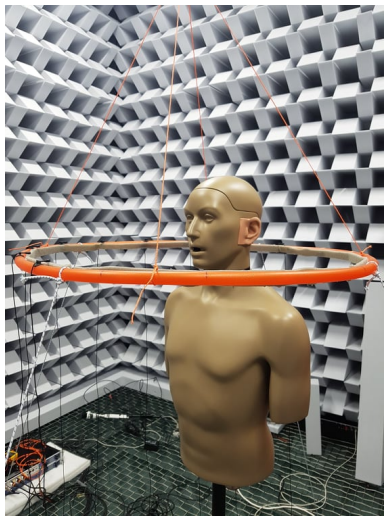


Fig. 3: KEMAR with mouth simulator inside the semi-cylindrical microphone array.

justed individually to each subject's mouth. Furthermore, subjects were instructed to stand still, although small head movements were expected and accepted as a part of natural speech.

The spoken content recorded for each subject was the first set of 10 phonetically balanced "HARVARD" sentences [32]. Subjects repeated the set of sentences four times at normal speaking levels. The reason for the repetitions was to allow subjects to become more familiar with the phrases, such that they could speak progressively more naturally. During the recordings, subjects also wore an in-ear monitor through which they were cued when to speak, as well which sentence of the set they should repeat. Once the sentence was spoken, the next sentence in the set would then be presented. For the first two repetitions of the set of sentences, the subjects were allowed to hold and read a piece of paper containing the set of phrases. For the last two repetitions, the paper was taken away such that the sentences were repeated solely after being cued by the in-ear monitor. The position of the subjects were checked between sets for possible misalignment, and corrected if necessary. From the four sets of repetitions, the last repetition was kept for it was usually the most natural of all four performances.

2.1.2 Mouth Simulator Measurements

The G.R.A.S. KEMAR 45BC head & torso with mouth simulator was placed in the middle of ring in a similar fashion to the human subjects, facing the 0° microphone with its mouth in the center of ring, as shown in Fig. 3.

The mouth simulator was equalized to a flat response at the position of the frontal microphone of the array using a G.R.A.S 46AF 1/2" free-field microphone calibrated to 94 dB at 1 kHz. The microphone was placed 36.5 cm away from the lips of the manikin instead of 2.5 cm, as recommended from the manual [21], in order to reconstruct the recording conditions of the speaking subjects. The level of the manikin was then set to 94 dB at 2.5 kHz, the minima

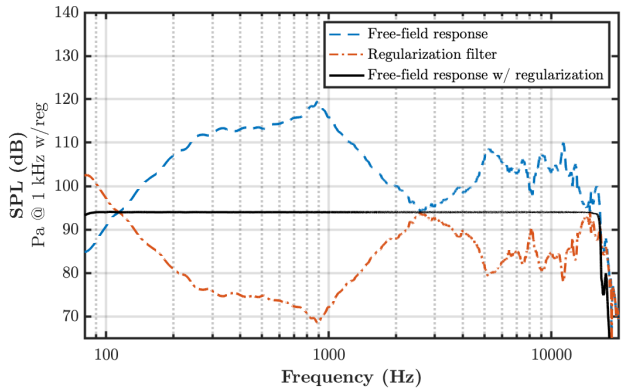


Fig. 4: The measured, inverted and resulting frequency response of the KEMAR 45BC mouth simulator equalization.

of the magnitude frequency response. For the equalization, Kirkeby regularization was applied with in- and out-band regularization parameters of 0 dB and 30 dB respectively, between 80 Hz and 13 kHz with no octave smoothing. Fig. 4 presents the measured, inverted and resulting equalized frequency response.

Impulse responses of the KEMAR mouth simulator were captured using the swept-sine technique between the equalized mouth and all 9 microphones of the array. Impulse responses were captured from the KEMAR, instead of directly reproducing speech through the device, to minimize the effects of noise on the further evaluation steps and improve its perceived directional quality.

2.1.3 Loudspeaker Measurements

A Genelec 8020B loudspeaker was also measured as a near-field speech source. The frontal baffle of the loudspeaker was positioned in the middle of the ring with the height of the ring adjusted to its acoustics center between the woofer and tweeter cones of the speaker. The loudspeaker was equalized using the same technique as for the KEMAR, and IRs were also measured in similar fashion to the KEMAR in the previous section.

2.2 Piston on Spherical Baffle

An analytical model previously used to emulate the sound radiation of human speech, is a piston on a rigid spherical baffle [8]. The rigid sphere is supposed to emulate the scattering effects of the head while the piston is supposed to emulate the flow of air through the mouth during speech. Assuming the geometry of the sphere and piston is axis-symmetric, the pressure around this model centered at the mouth can be described by:

$$p(r,\phi) = \frac{i\rho_0 cW}{2} \sum_{n=0}^{\infty} \left[P_{n-1}(\cos\alpha) - P_{n+1}(\cos\alpha) \right] \frac{h_n(kr_c(r))}{h_n'(kR)} P_n(\cos\phi'(\phi)) \quad (1)$$

where i is the imaginary number, k is the wavenumber, ρ_0 is the density of air, c is the speed of sound, and W is the

piston's velocity in the normal direction. Furthermore, R is the radius of the sphere, and α is the radius of the vibrating piston. Since the origin is centered at the mouth instead of the head, r_c is the distance from the center of the sphere as a function of r, the distance from the mouth, or

$$r_c(r) = \sqrt{R^2 + r^2 + 2Rr\cos\phi'}, \tag{2}$$

while ϕ' is the angle from the frontal axis centered at the head as a function of ϕ, the angle from the frontal axis centered at the mouth, or

$$\phi'(\phi) = \begin{cases} \arcsin\left(\frac{r}{r_c}\sin\left(\pi-\phi\right)\right), & \text{for } \phi = [-\frac{\pi}{2}, \frac{\pi}{2}] \\ \pi - \arcsin\left(\frac{r}{r_c}\sin\left(\pi-\phi\right)\right), & \text{for } \phi = (\frac{\pi}{2}, \frac{3\pi}{2}) \end{cases}. \tag{3}$$

Finally, P_n and h_n refer to Legendre polynomials of degree n and spherical Hankel functions of order n, with $(')$ indicating its derivative, The radius R for the sphere was defined as 95 mm and the radius α for the piston as 9 mm as previously implemented by [6]. The implementation of this methods was verified by comparing and matching the results presented in [22].

2.3 Circular Harmonic Decomposition

The 2-dimensional directivity pattern for some plane of a source can be expressed as a Fourier series representation:

$$D(\phi) = \sum_{m=-\infty}^{\infty} \gamma_m \, \Phi_m(\phi) \tag{4}$$

where Φ_m corresponds to the orthonormal set of circular basis functions of order m or

$$\Phi_m(\phi) = \frac{1}{\sqrt{2}} \begin{cases} \sqrt{2}\cos(m\phi), & \text{for } m > 0 \\ 1, & \text{for } m = 0 \\ \sqrt{2}\sin(|m|\phi), & \text{for } m < 0 \end{cases} \tag{5}$$

and γ_m are the Fourier coefficients of the decomposed directivity [17]. In practice, this expansion is limited to an order M which minimizes truncation error and spatial aliasing, where $2M+1$ sampling points are required to reconstruct basis functions of order M.

Directivity patterns produced from speech averages, as described in Section 3, can be fitted to a set of circular harmonic basis functions providing there are sufficient sampling points. As the directivity captured by the array of microphones was assumed to be symmetric along the forward facing direction, the pressure of 7 microphones of the array were mirrored along the x-axis defining the pressure over an entire circle and allowing for a 7^{th} order CH decomposition.

The computation of the Fourier coefficients γ from the discrete sampled directivity measurements can be performed through the least-squares solution [17] to:

$$\mathbf{d}(k) = \mathbf{\Phi}\,\boldsymbol{\gamma}(k) \tag{6}$$

where $\mathbf{d}$ is now a vector of a size $Q \geq 2M+1$, Q is the number of sample points, containing the frequency dependent directivity values, $\mathbf{\Phi}$ is a $Q \times (2M+1)$ matrix containing the circular harmonic basis functions at the position

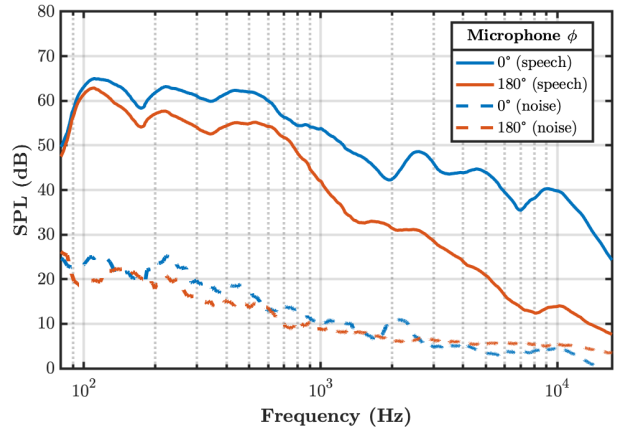


Fig. 5: Long-term average frequency spectrum with ⅓ octave smoothing for speech and microphone self-noise for a single subject.

of the sensors, and $\boldsymbol{\gamma}$ is a vector $2M+1$ containing Fourier coefficients up to order M.

2.4 Directivity Implementation

To emulate the directivity effects of the various devices and methods on speech, the recorded speech for the frontal direction of the 24 subjects was convolved with the measured IRs of the KEMAR and loudspeaker, as well as the filters derived from the PoS model and CH decomposition, for the other 8 directions.

The long-term average spectrum and microphone noise floor for the speech recording of a subject is presented in Fig. 5. For rear-incident directions, the noise became slightly audible because the signal-to-noise ratio at high frequencies decreases due to the acoustic shadowing effects of the head. The process of capturing IRs for the KEMAR and loudspeaker reduced the measured level of the noise produced by these devices as well as the inherent microphone noise. Furthermore, the filters derived from the model and decomposition inherently do not contain noise. To match the noise quality between the recorded speech and the directivity-synthesized speech from the devices and sources, portions of recorded silence containing noise from each microphone of the array were added to the corresponding speech directions. The spectrum and RMS of the added noise matched that of the original recordings.

3 EVALUATION

The repeatable speech sources were evaluated through objective and subjective comparisons with the speech from recorded subjects. The metrics used for objective comparisons were speech directivity and spectral differences. For subjective comparisons, a MUSHRA listening test was performed.

Though speech signals contain a variety of noise-like, impulsive, and harmonic components, and the spectrum of speech varies rapidly and dynamically with time, it is common to analyse speech directivity through the average of its

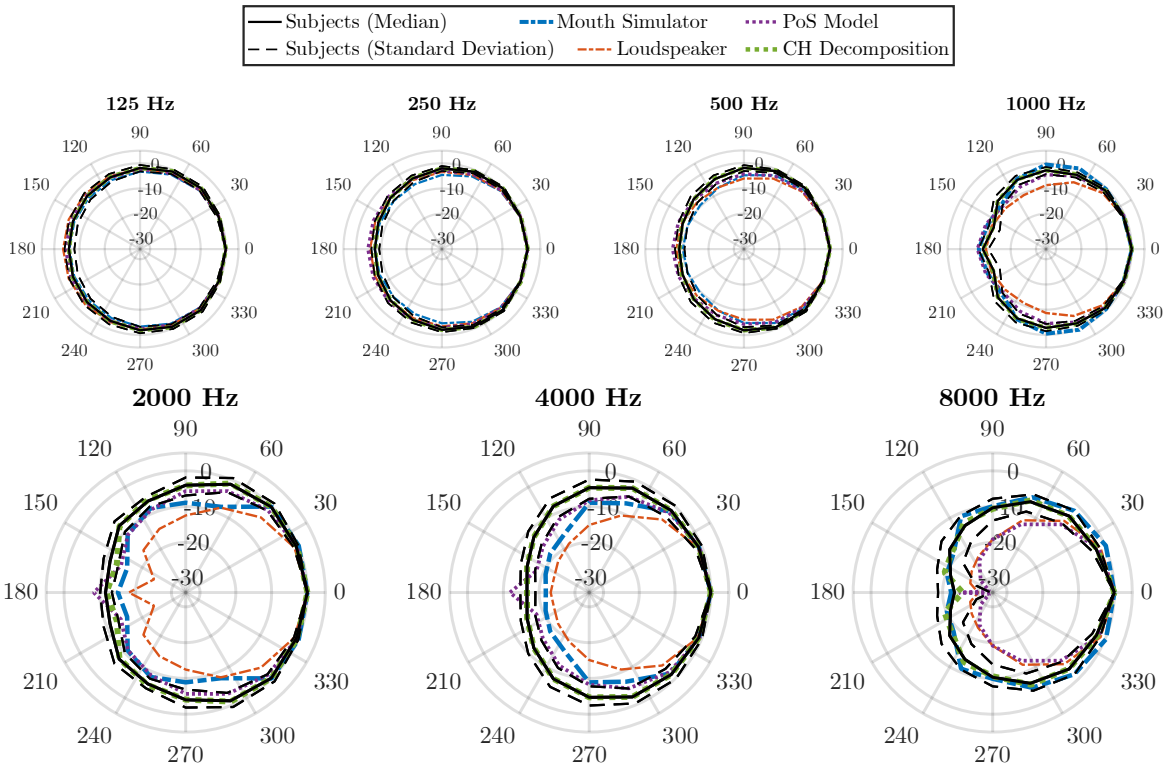


Fig. 6: Speech directivity comparison between the different tested repeatable speech sources, where PoS and CH denote Piston on Sphere and Circular Harmonic, respectively.

magnitude spectrum. The recorded speech was processed in overlapping frames of 1 second at a sampling frequency of 48 kHz with a hop size of 3/4 of a frame. To avoid processing significant pauses, an equivalent sound level (L_{eq}) of the frame was calculated as:

$$L_{\text{eq}} = 20\log\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{p[t]}{p_{\text{ref}}}\right)^2\right] \tag{7}$$

where p corresponds to the pressure of the speech signal and p_{ref} is the reference sound pressure of 20^{-5} pascals. If the L_{eq} dropped below a 50 dB threshold, a pause was assumed and the frame was omitted from processing. If no pause was assumed, a Hann window was applied to the frame and then transformed into the frequency domain. Finally, the magnitude of all speech windows was averaged and analyzed by frequency.

3.1 Directivity Patterns

A well established method for analyzing the radiation of speech is that of its directivity pattern. This directivity pattern is defined in decibels D_{dB} as the ratio between the average frequency magnitude for sound radiated in a direction of interest and the sound radiated in the frontal directions, or:

$$D_{\text{dB}}(\phi) = 20\log_{10}\left(\frac{|H(\phi)|}{|H(0^\circ)|}\right) \tag{8}$$

where H corresponds to the frequency domain transform of an IR or the average magnitude of speech spectra, and ϕ is the angle from the frontal direction.

The average directivity pattern from Eq. (8) was obtained for the recorded subjects, KEMAR, loudspeaker, PoS and CH model. Before estimating the directivities in decibels, the directivities in the linear scale were averaged in 1/3 octave wide bands for 7 center frequencies. Fig. 6 graphically presents these directivities in a polar form to communicate the spatial distribution of the radiating source. Because the directivity of speech is assumed to be symmetrical, the presented polar plots mirror the directivity along frontal direction.

An expected trend is observed among the directivity patterns. At low frequencies, the directivities of the sources tend toward omnidirectional due to the size of the produced wavelength being larger than the size of the source. As the frequency increases and the wavelength becomes proportional to the size of the radiating element (average mouth opening, loudspeaker driver, piston), the scattering of the source's body becomes more prominent and the sound field radiated by the source becomes more directive. In regards to the speech directivity results, -6 to -20 dB differences are seen between front and rear directions from 250 Hz to 8 kHz respectively, and -3 to -12 dB for lateral directions. Also, at 1 kHz, a small back lobe can be observed.

3.2 Spectral Magnitude Difference

Another way to compare reproducible speech sources is the difference between their spectral magnitude for various radiating directions. In the case of speech sources, this difference is calculated between the recorded speech, which is assumed as the reference, and a source of comparison.

The spectral magnitude differences are then estimated in decibels by:

$$\textbf{coloration}(\phi) = 20\log_{10}\left(\frac{|H_{\text{KEMAR}}(\phi)|}{|H_{\text{Subjects}}(\theta)|}\right) \quad (9)$$

The average spectral coloration was evaluated for the four tested repeatable speech sources with respect to the speech recorded from human subjects. Fig. 7 presents the coloration results for all recorded angles.

All tested speech sources produced minimal coloration below 800 Hz. The mouth simulator produced relatively low spectral magnitude differences in general, though with sharp peaks around 9 kHz at lateral directions as well as high frequency roll off at rear directions. The coloration produced by the loudspeaker was mainly notches that were wide in bandwidth at some directions, and occurred most prominently at the lateral directions. The PoS model produced relatively similar magnitude responses to the measured human speakers at mid-high frequencies, with the exception of the rear direction, however with the most pronounced high frequency roll-off. The CH decomposition produced the least colored responses overall.

3.3 Subjective Evaluation

Finally, in order to subjectively evaluate the KEMAR mouth simulator as a perceptually-plausible human person speaking in the near-field, a MUSHRA listening test [33] was performed.

3.3.1 Test Paradigm

The reference stimuli were samples from the recorded male and female speech from Section 2.1.1, as these recordings were assumed to contain the natural directional sound qualities of human speakers. For each MUSHRA slide, the subjects were required to compare the reference with the speech reproductions from the KEMAR mouth simulator, the loudspeaker, as well as speech synthesized with the directivity of the PoS model and CH decomposition filters, for all nine recorded radiation angles. After two training slides, each radiation angle was tested 3 times so the entire test involved 27 MUSHRA slides. As specified in Section 2.4, noise from the microphones of the original speech recordings were added to the other stimuli to match the quality of the reference recordings. Subjects were asked to rate the perceived coloration, the dissimilarity in timbre between stimuli [34], for various stimuli on a continuous scale from 0 to 100 with respect to the reference. A verbal scale was included with the terms: **Same**, **Similar**, **Somewhat Similar**, **Somewhat Dissimilar**, and **Dissimilar**, next to the continuous scale at scores 10, 30, 50, 70, and 90, respectively. The stimuli presented to the listeners was monophonic in format, as opposed to binaural, to prevent adding potential extra coloration as well as wrong localization cues due to non-individualized HRTF mismatches. Amongst the stimuli, there was also a hidden reference and anchor, which was produced by applying a 3.5 kHz cut-off low pass filter to the reference. The A-weighted RMS between all stimuli was matched individually for each direction. The listening test was performed in quiet listening booths with Sennheiser HD 650 head-

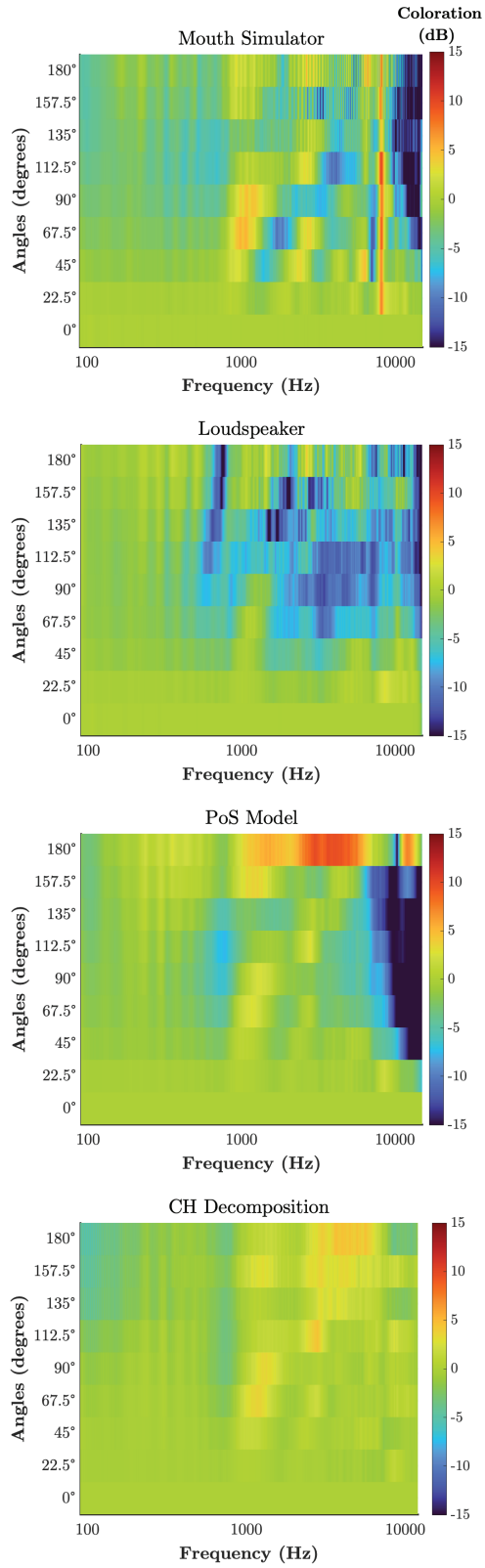


Fig. 7: Average coloration between the speech of 24 subjects and the speech reproduced by the mouth simulator, loudspeaker, piston on a sphere model and circular harmonic decomposition filters for various angles.

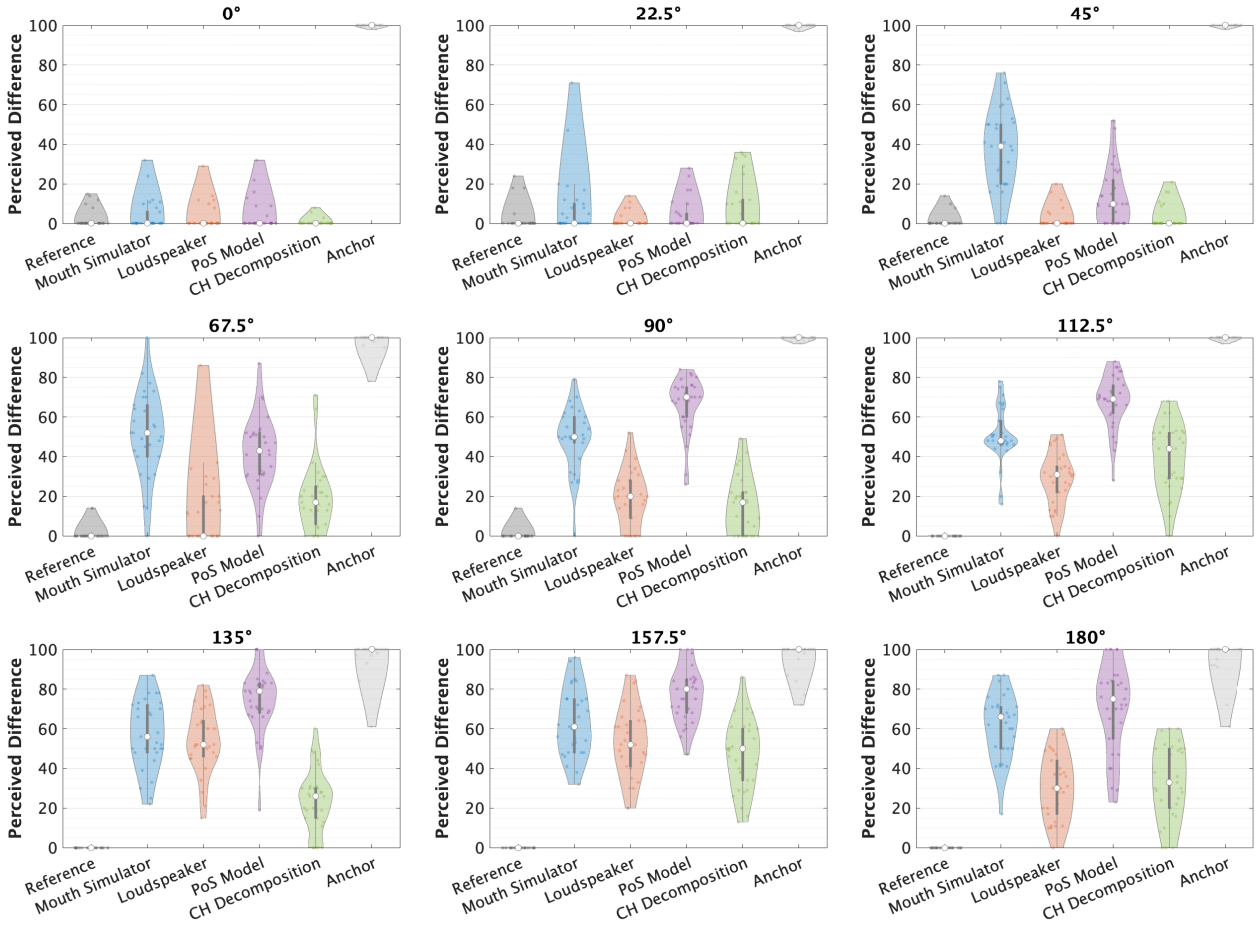


Fig. 8: Violin plots of the listening test results for all tested angles and repeatable speech sources, where PoS and CH denote Piston on Sphere and Circular Harmonic, respectively.

phones, calibrated following the ITU-R BS.1534-3 standard [33]. Stimuli and conditions were randomised and presented double blind, and each tested direction was repeated three times.

3.3.2 Results

Twelve participants aged between 19 and 37 (2 female and 10 male) took part in the experiment, with self reported normal hearing and prior critical listening experience (such as education or employment in audio or music engineering).

The results of two participants were excluded from analysis due to consistently high ratings of the anchor, as recommended in [33]. The results of the listening test are presented as violin plots [35] in Fig. 8 for all tested angles. Violin plots display the density trace and box plot together, which better illustrates the structure of the data than traditional box plots. The data was tested for normality using the Shapiro-Wilk test. Even excluding the reference and anchor data, not all data was shown to be normally distributed. Therefore non-parametric statistical analysis was conducted.

Friedman tests were conducted to assess whether there were statistically significant differences between the ratings of different stimuli, with results for the anchor excluded. The tests were highly statistically significant for all angles ($p < 0.001$) except $\phi = 0°$: $\chi^2(4) = 5.99, p = .199$ and $\phi = 22.5°$: $\chi^2(4) = 15.7, p = 0.004$, though the significance at $\phi = 22.5°$ falls comfortably within a 95% confidence interval. Note that due to the equalization of the repeatable speech sources to $\phi = 0°$, no significant differences at frontal angles were expected.

Looking at the results for the different tested repeatable speech sources, it is clear the Mouth Simulator was rated as considerably different from the Reference for most tested angles, the Loudspeaker and CH Decomposition were the closest rated to the reference, and the PoS Model was perceived as the most different for 5 out of the 9 tested angles. To test the statistical significance, post-hoc Wilcoxon signed rank tests using the Bonferroni-Holm correction were conducted; the results of which are presented in Fig. 9. At $\phi =\geq 90°$, none of the tested repeatable speech sources were indistinguishable from the recorded speech ($p > 0.05$). At frontal incidences however, the Loudspeaker produced statistically significantly similar rendering to the recorded speech at a confidence interval of 95% ($p < 0.05$).

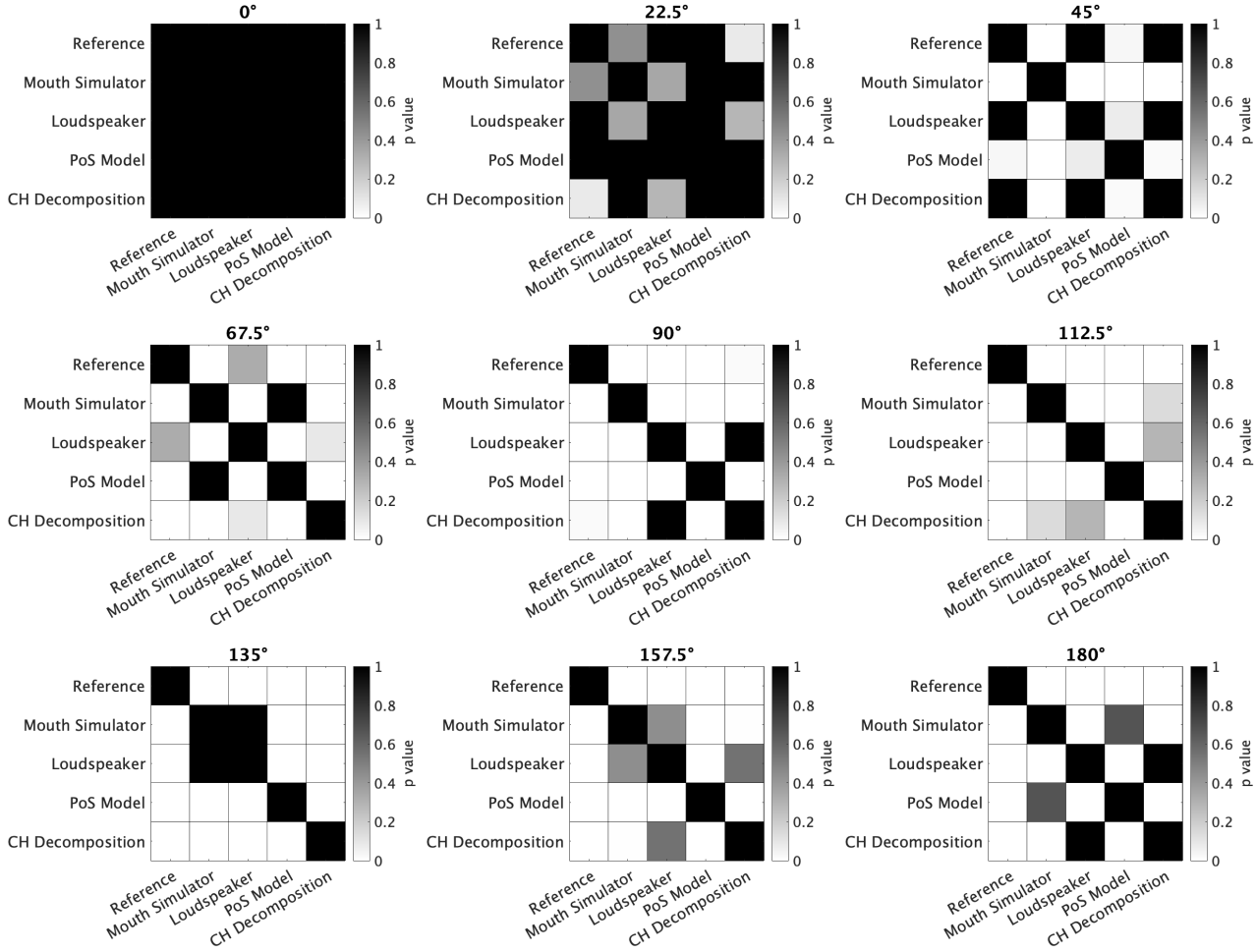


Fig. 9: Wilcoxon signed rank (with the Bonferroni-Holm correction) matrix plots of the listening test results for all tested angles and repeatable speech sources, where PoS and CH denote Piston on Sphere and Circular Harmonic, respectively.

4 DISCUSSION

The directivity patterns shown in Fig. 6 are in general as expected. These directivity results follow similar trends as those presented in previous far-field studies of speech directivity [7, 9, 11], though further work would be required to evaluate the differences between repeatable near- and far-field speech reproduction.

All evaluated sources emulate speech directivity well below 1 kHz, above which variations are seen for the mouth simulator, KEMAR and PoS model. The directivity of the mouth simulator differs from speech for lateral directions either by more energy at 1 kHz, or less energy at 2 and 4 kHz. This differs from the findings presented in [6], which found mouth simulators to be more directive than human speech at high frequencies. These differences may be because those measurements were conducted at 1 m, in the far-field. The mouth simulator also produces less energy at rear directions at 4 kHz. From 1 kHz onwards, the loudspeaker displays less energy and is more directive than speech for all off-axis directions. Besides its backwards radiation at 2 kHz, the directivity of the loudspeaker is smooth for all frequencies. The PoS model follows the trend of speech directivity for all directions until about 4 kHz, from which it becomes increasingly more directive with the exception of a narrow but strong back lobe at 8 kHz. As expected, the directivity of the CH decomposition closely follows the average speech directivity, as the model is derived from data from on the latter.

The spectral magnitude difference results from Fig. 7 indicate minimal coloration for frontal directions, which is to be expected due to the equalization to the frontal direction. Timbral mismatches between the reference and other sources become increasingly significant towards lateral and rear directions. For lateral directions, the mouth simulator shows more energy at 1 kHz and less energy at 2 kHz, results which match the directivity results of Fig. 6. The strong coloration around 9 kHz, which seems unaccounted in the directivity plots, may be an artefact of the frontal-only equalization. A diffuse-field equalization approach may improve this, though at the expense of greater frontal coloration. The loudspeaker showed strong coloration at the lateral directions, which is reflected by the narrow directivity observed in the directivity plots (see again Fig. 6). The strong coloration of the PoS model at the rear direction is likely explained by the large back lobe in the directivity,

whereas the CH decomposition was the least colored of the four tested methods.

Results for the subjective evaluation presented in Fig. 8 indicate that listeners were not able to discern major differences between the reference and the rest of the sources for the two tested frontal directions. This observation is supported by the p-values for $\phi = 0°$ and 22.5° which are larger than .001, failing to reject the null-hypothesis and suggesting no significant perceptual differences between stimuli, though the $\phi = 22.5°$ direction was significant at a confidence interval of 95%. This is expected, considering the re-recorded speech samples were equalized to the frontal microphone of the human speech samples.

However, for lateral and rear directions, the results show that listeners were able to discern between reference speech and other sources. Starting from 45°, p-values below .001 indicate highly-significant perceptual differences. Still, listeners gave alternative sources mid-high ratings, such as **Somewhat Similar**, on the verbal scale, which suggests that their speech reproduction quality was not necessarily poor. These listening test results correlate with the trends presented in Fig. 7, supporting the hypothesis that coloration affects the quality of perceived speech, and are not unexpected considering human sensitivity to timbral variances in speech perception [36, 37]. The combination of Fig. 7 and Fig. 8 suggests that listeners were less sensitive to negative coloration, i.e. notches in the frequency response, as opposed positive coloration, i.e. peaks in the frequency response, something that has been reported before [38].

In summary, the circular harmonic decomposition appears to be an efficient method, performing well in the tested metrics. However, as it requires extra technology for real-time implementation, such as low latency head-tracked dynamic binaural rendering, it is not the most practical option. Based on the results of the subjective evaluation, an approximately head-sized high-quality monitor loudspeaker such as the tested Genelec model may be a reasonable choice for a speech source in the absence of a proper mouth simulator. Further in-situ tests should be conducted to confirm this, though.

5 CONCLUSION

This study has evaluated the near-field directivity accuracy of four methods for reproducible speech, which could be beneficial for immersive telepresence applications, inducing a higher level of intimacy and immersion. The methods evaluated have included two physical; a mouth simulator and traditional loudspeaker, and two numerical; a piston on a sphere model and a circular harmonic decomposition.

Near-field speech directivities have been captured for 24 human subjects, and compared with the directivities of the four repeatable speech methods both objectively, through directivity patterns and spectral magnitude difference calculations, and perceptually, through a listening test on perceived coloration. The results have shown that, while there are few perceived differences for frontal directions between reference speech and properly equalized sources, sources with non-smooth directivities can introduce audible coloration into directional speech, spectral peaks in lateral and rear directions are perceived as more colored than spectral notches. The closest near-field directivities were produced by the circular harmonic decomposition method and mouth simulator, however the least coloration was perceived in the circular harmonic decomposition method and the loudspeaker.

The measurements and renders used in this study were obtained at a single radius of 36.5 cm. Future work could conduct perceptual comparisons of the repeatable speech sources between the near-field and far-field, at distances starting from the extreme near-field (speaking just in front of the ear).

Finally, the repeatable speech sources evaluated in this study were all perceived with statistically significantly greater coloration than the recorded human speech at angles greater than 90°. However, the extent to which the non-authentic off-axis speech reproduction would affect realistic application scenarios, such as in-situ augmented reality scenarios with both real and virtual speech sources, remains unknown. Future work could assess the plausibility of repeatable speech to determine the 'minimum requirement' threshold of reproduction accuracy.

6 ACKNOWLEDGEMENTS

The authors would like to thank the members of the Dominante choir for participating as speech subjects. Thomas McKenzie was supported by the Human Optimised XR (HumOR) Project.

7 REFERENCES

[1] J. Steuer "Defining Virtual reality: Dimensions Determining Telepresence," *J. Comm.*, vol. 42, no. 4, pp. 73-93 (1992), https://doi.org/10.1111/j.1460-2466.1992.tb00812 .

[2] D. Potdevin, C. Clavel, N. Sabouret. "Virtual Intimacy, This Little Something Between Us," in *Proc. 18th International Conference on Intelligent Virtual Agents* (Sidney, Australia) (2018), https://doi.org/10.1145/3267851.3267884.

[3] J. P. Noel, A. Serino, M. T. Wallace "Increased Neural Strength and Reliability to Audiovisual Stimuli at the Boundary of Peripersonal Space," *J. Cogn. Neurosci.* 31 vol. 8, pp. 1155–1172 (2019), https://doi.org/10.1162/jocn_a_01334.

[4] A. Batliner, B. Schuller, S. Schaeffler and S. Steidl, "Mothers, Adults, Children, Pets — Towards the Acoustics of Intimacy," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* pp. 4497-4500 (2008), https://doi.org/10.1109/ICASSP.2008.4518655.

[5] M. Brandner, R. Blandin, M. Frank, and A. Sontacchi "A Pilot Study on the Influence of Mouth Configuration and Torso on Singing Voice Directivity," *J. Acoust. Soc. Am.*, vol. 148, pp. 1169 (2020), https://doi.org/10.1121/10.0001736.

[6] T. Halkosaari, M. Valgamaa, and M. Karjalainen "Directivity of Artificial and Human Speech," *J. Audio Eng. Soc.*, vol. 53, no. 7/8, pp. 620–631 (2005).

[7] H. K. Dunn and D. W. Farnsworth "Exploration of Pressure Field Around the Human Head During Speech", *J. Acoust. Soc. Am.*, vol. 10, no. 1, pp. 184-199 (1939), https://doi.org/10.1121/1.1915975.

[8] J. L. Flanagan "Analog Measurements of Sound Radiation from the Mouth," *J. Acoust. Soc. Am.*, vol. 32, no. 12, pp. 1613-1621 (1960), https://doi.org/10.1121/1.1907972.

[9] A. Chu and A. Warnock "Detailed Directivity of Sound Fields Around Human Talkers," Report B3144.6 prepared for Public Works and Government Services Canada, (Sep 2001), https://doi.org/10.4224/20378930.

[10] B. F. G. Katz, F. Prezat, and C. d'Alessandro "Human Voice Phoneme Directivity Pattern Measurements," *J. Acoust. Soc. Am.*, vol. 120, pp. 3359 (2006), https://doi.org/10.1121/1.4781486.

[11] B. B. Monson and E. J. Hunter "Horizontal Directivity of Low- and High-Frequency Energy in Speech and Singing," *J. Acoust. Soc. Am.*, vol. 132, no. 1, (2012), https://doi.org/10.1121/1.4725963.

[12] P. Kocon and B. B. Monson "Horizontal Directivity Patterns Differ Between Vowels Extracted from Running Speech," *J. Acoust. Soc. Am.*, vol. 144, no. 1, (2018), https://doi.org/10.1121/1.5044508.

[13] C. Pörschmann and J. M. Arend "Investigating Phoneme-Dependencies of Spherical Voice Directivity Patterns," *J. Acoust. Soc. Am.*, vol. 149, pp. 4553-4564 (2021), https://doi.org/10.1121/10.0005401.

[14] T. W. Leishman, S. D. Bellows, C. M. Pincock, and J. K. Whiting "High-resolution Spherical Directivity of Live Speech from a Multiple-Capture Transfer Function Method," *J. Acoust. Soc. Am.*, vol. 149, pp. 507–1523 (2021), https://doi.org/10.1121/10.0003363.

[15] C. Pörschmann and J. M. Arend "A Method for Spatial Upsampling of Voice Directivity by Directional Equalization," *J. Audio Eng. Soc.*, vol. 68, no. 9, pp. 649-663 (2020), https://doi.org/10.17743/jaes.2020.0033.

[16] H. F. Olson "Field-Type Artificial Voice," *J. Audio Eng. Soc.*, vol. 20, no. 6, pp. 446-452 (1972).

[17] F. Zotter and M. Frank "Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality" pp. 58-59 (Springer Open, 2019), https://doi.org/10.1007/978-3-030-17207-7.

[18] ITU-T Rec. P.58 *Head and Torso Simulator for Telemetry* (2021).

[19] F. Bozzoli, M. Viktorovich, and A. Farina "Balloons of Directivity of Real and Artificial Mouth Used in Determining Speech Transmission Index," in *Proc. AES 118th Convention* pp. 1-5 (Barcelona, Spain) (2005).

[20] D. Cabrera, H. Sato, W. L. Martens, and D. Lee "Binaural Measurement and Simulation of the Room Acoustical Response from a Person's Mouth to Their Ears," *Australian Acoust. Soc.* vol. 37, no. 3, pp. 98-103 (2009), .

[21] G.R.A.S Sound & Vibration: "Instruction manual: G.R.A.S. KEMAR 45BC". (2013) available at www.grasacoustics.com/files/m/a/man_45BB_45BC.pdf.

[22] R. M. Aarts, A. J. M. Janssen "Sound Radiation from a Resilient Spherical cap on a Rigid Sphere," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2262-2273 (2010), https://doi.org/10.1121/1.3303978.

[23] G. Fischer, C. Schneiderwind, and A. Neidhardt "Comparing the Directivity of a Mouth Simulator and a Simple Physical Model," in *45th Annual Meeting on Acoustics (DAGA)* (Rostock, Germany) (2019), .

[24] T. McKenzie "Assessing the Authenticity of the KEMAR Mouth Simulator as a Repeatable Speech Source," in *Proc. AES 143rd Convention* (Milan, Italy) (2017).

[25] E. G. Williams "Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography" (Academic Press, 1999), https://doi.org/10.1016/B978-0-12-753960-7.X5000-1.

[26] F. Zotter, *Analysis and Synthesis of Sound-Radiation with Spherical Arrays*, Ph.D. thesis, Institute of Electronic Music and Acoustics University of Music and Performing Arts, Austria (2009).

[27] D. S. Brungart & W. M. Rabinowitz "Auditory Localization of Nearby Sources. Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1465–1479 (1999), https://doi.org/10.1121/1.427180.

[28] K. Young, C. Armstrong, A. I. Tew, D. T. Murphy, G. & Kearney "A Numerical Study into Perceptually-Weighted Spectral Differences Between Differently-Spaced HRTFs," in *AES International Conference on Immersive and Interactive Audio*, pp. 1-10 (York, UK) (2019), .

[29] www.grasacoustics.com/products/measurement-microphone-sets/product/145-46af.

[30] A. Farina "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *Proc. AES 108th Convention*, pp. 1-23 (Paris, France) (2000).

[31] O. Kirkeby, P. A. Nelson, and F. Orduna-Bustamante "Fast Deconvolution of Multichannel Systems using Regularization," *IEEE Trans. Speech and Audio Process.*, vol. 6, no. 2, pp. 189–194 (1998), https://doi.org/10.1109/89.661479.

[32] IEEE Subcommittee on Subjective Measurements "IEEE Recommended Practices for Speech Quality Measurements," *IEEE Trans. Audio and Electroacoustics*, vol. 17, no. 297, pp. 227–246 (1969), https://doi.org/10.1109/TAU.1969.1162058.

[33] ITU-R BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems, (2015).

[34] ITU-R BS.2399-0: Methods for Selecting and Describing Attributes and Terms,in the Preparation of Subjective Tests, (2017).

[35] J. L. Hintze and R. D. Nelson "Violin Plots: A BoxPlot-Density Trace Synergism," *The Am. Statistician*, vol. 52, no. 2, pp. 181–184 (1998).

[36] D. H. Klatt and L. C. Klatt "Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers," *J. Acoust. Soc. Am.*, vol. 87, pp. 820-857 (1990), https://doi.org/10.1121/1.398894.

[37] C. E. Stilp and A. A. Assgari "Perceptual Sensitivity to Spectral Properties of Earlier Sounds During Speech Categorization," *Attention, Perception, & Psychophysics.*, vol. 80, pp. 1300–1310 (2018), https://doi.org/10.3758/s13414-018-1488-9.

[38] R. Bücklein "Audibility of Frequency Response Irregularities," *J. Audio Eng. Soc.*, vol. 29, no. 3, pp. 126–131 (1981).