Investigation of the influence of the torso, lips and vocal tract configuration on speech directivity using measurements from a custom head and torso simulator

Rémi Blandin* Jingyan Geng, and Peter Birkholz

Institute of Acoustics and Speech Communication, TU Dresden, Dresden 01062, Germany

Received 24 October 2022, Accepted 6 July 2023

Abstract – The human voice is a directional sound source. This property has been explored for more than 200 years, mainly using measurements of human participants. Some efforts have been made to understand the anatomical parameters that influence speech directivity, e.g., the mouth opening, diffraction and reflections due to the head and torso, the lips and the vocal tract. However, these parameters have mostly been studied separately, without being integrated into a complete model or replica. The aim of this work was to study the combined influence of the torso, the lips and the vocal tract geometry on speech directivity. For this purpose, a simplified head and torso simulator was built; this simulator made it possible to vary these parameters independently. It consisted of two spheres representing the head and the torso into which vocal tract replicas with or without lips could be inserted. The directivity patterns were measured in an anechoic room with a turntable and a microphone that could be placed at different angular positions. Different effects such as torso diffraction and reflections, the correlation of the mouth dimensions with directionality, the higher-order modes and the increase in directionality due to the lips were confirmed and further documented. Interactions between the different parameters were found. It was observed that torso diffraction and reflections were enhanced by the presence of the lips, that they could be modified or masked by the effect of higher-order modes and that the lips tend to attenuate the effect of higher-order modes.

Keywords: Directivity, Speech, Experiment, Replica, Vocal tract

1 Introduction

1.1 General context

The human voice is a directional sound source whose amplitude is highest towards the front of a person and progressively decreases toward the sides and back of a person. This was documented more than 200 years ago [1], and it has been documented with increasing accuracy over the past few decades [2–8]. The directionality and the shape of the directivity patterns depend on the phoneme and the articulation and anatomy of the participants. Knowledge about speech directivity is of interest for various applications, such as the acoustic design of rooms, microphone placement, telephony and auralization in three-dimensional (3D) environments. Studying how speech directivity is influenced by various parameters such as the vocal tract shape, the lips or the torso is useful for identifying the most influential parameters and improving the accuracy of speech radiation modeling. These parameters have been explored through simulations and experiments with replicas [9–15].

1.2 Speech radiation models

The simplest theoretical model for speech directivity is a piston placed on the surface of a sphere [9, 11]. It describes the diffraction caused by the head and the influence of the mouth area. It predicts that the amplitude decreases toward the sides of and behind the sphere, and the directionality increases as the frequency and the mouth area increase. These features are in agreement with the trends observed in measurements of human speakers. However, for human speakers, at very high frequencies (starting at about 10 kHz), the directionality stops increasing or even decreases [16]. The effect of the mouth area was observed only in the 2 kHz and 4 kHz octave bands [3, 8, 16, 17]. On the other hand, the predicted shape of the radiation pattern is much simpler than the patterns measured using human participants. An improved version of this model uses a prolate spheroid to better approximate the head geometry and describes very roughly the reflections on the torso with an infinite baffle [10]. This seems to improve the agreement with measurements of human participants; however, very few comparisons with this improved model have been made.
More accurate models based on numerical methods, such as finite element methods, and a more accurate description of the head shape, including the vocal tract and part of the torso, have been developed. However, to our knowledge, these models were very rarely used to predict speech radiation patterns; instead, they were mainly used to obtain a better description of the boundary condition at the mouth opening for the computation of vocal tract transfer functions. One can note, however, the work of Kagawa et al. [18], who predicted the acoustic field at some specific frequencies in the vertical and horizontal planes. These data highlight some complex patterns not predicted by the simpler models, but they are difficult to compare with measurements of humans, which are mostly provided as angle-dependent relative amplitudes; they are not provided in this work. More recently, finite element simulations were used to investigate the impact of the accuracy of the geometrical description of the head and lips on the computation of the vocal tract transfer function [12]. It was found that the transfer function was mainly affected by the presence of the lips and that the other details of the head had a limited influence. A model consisting of a sphere with lips gave very similar results to a realistic head model. However, this study did not explore the impact of such geometrical simplifications on the directivity. Birkholz et al. [19] also used a finite element model to predict the transfer function from the mouth exit to a point 30 cm in front of the mouth in order to study how the accuracy of the radiation modeling impacts the naturalness and intelligibility of articulatory synthesis. Substantial deviations between the radiation characteristics computed with the simplest model and the radiation characteristics computed with the finite element simulation were found, highlighting the effect of the torso. However, since the focus was on only one direction, no radiation pattern was extracted from the simulations. Finally, it is worth mentioning that a boundary element model of a realistic head and torso, including magnetic resonance image (MRI)-based vocal tract geometries, was built recently to specifically study speech radiation [20]. This showed that the vocal tract affects the direction of radiation and the shape of the directivity patterns at high frequencies. However, this work is preliminary and would need to be pushed further to strengthen its conclusions.

Some works have specifically focused on the effect of higher-order modes on speech radiation. This was motivated by the basic principle of the multimodal method, which is to decompose the sound field inside the vocal tract into the transverse propagation modes [21]. Since higher-order modes are explicitly described with this method, it is easy to predict and simulate their impact on the radiation, which is referred to hereafter as the higher-order-modes effect (HOME). It was found that they affect speech radiation at high frequencies (about 3–4 kHz and higher) by generating complex patterns that substantially vary within small frequency intervals (on the order of 100 Hz) [22]. It was predicted that this effect is stronger for open vowels (e.g., /a/) than for closed vowels (e.g., /u/) [14]. Similar patterns were found in measurements of human participants [15, 16, 23]. However, the state-of-the-art radiation models based on the multimodal method focus only on the vocal tract shape and neglect all the other parameters, such as the head, torso and lips.

1.3 Simulation of speech radiation with mechanical replicas

Speech radiation can be simulated with a head and torso simulator (HATS), which is a mannequin with average human dimensions. Such simulators are mostly used for binaural recordings to simulate head-related transfer functions (HRTF). However, a HATS equipped with a sound source to simulate the sound radiated by a human speaker can be used for some specific applications, such as telephone design or testing car acoustics. Some studies have compared the radiation characteristics of HATS with measurements of human participants [24, 25]. Some substantial differences have been found; they have been attributed to the simplification of the geometry (head and mouth) and the inability of the HATS to simulate different phonemes. Halkosaari et al. [24] proposed some improvements regarding the mouth aperture size and the filtering of the radiated sound for a specific position for telephone design.

A commercial HATS was used by Brandner et al. [15] to investigate the influence of the torso on the sound radiation. A comparison of the directivity patterns produced by the HATS with and without a torso showed that the torso generates an interference pattern with side lobes in the horizontal plane.

Some vocal tract replicas have also been used in isolation to study the HOME experimentally. The measured patterns showed a good agreement with the theoretical predictions [22]. Similar replicas, whose lips could be removed, were used to study the influence of the lips on the production and radiation of the consonant /s/ [13]. It was observed that the lips increase the directionality at high frequencies above 5 kHz. In both studies, the head was simulated with a flat baffle and no torso was simulated.

1.4 Objective

Theoretical and experimental explorations of the mechanisms influencing speech directivity have provided very valuable knowledge. However, so far, the different parameters (e.g., torso, lips, vocal tract) have been explored mostly in isolation from each other. For example, the impact of the torso has been evaluated independently, without considering the vocal tract and the lips. Likewise, the effects of the vocal tract and the lips have been studied by using a flat baffle to simulate the head, without considering the diffraction and reflections from the head and torso. This makes it impossible to evaluate their influence in combination. Hence, one could not identify potential interactions between parameters to, e.g., determine whether the effect of the torso is the same (preserved) if lips are present.

The objective of the present work is to investigate the effect of the vocal tract, the lips and the torso using a HATS that integrates all of these elements.
2 Methods

For this purpose, we designed a simplified HATS that consists of two spheres that are used to represent the head and the torso. This was inspired by the work of Algazi et al. [26], who used a similar design to study the HRTF. It was shown that such a simplified HATS can successfully reproduce the main features of the HRTF. Furthermore, such a design is simple to build, and it can be compared to fast multipole acoustic simulations [27]. Vocal tract replicas corresponding to the vowels /a/, /i/ and /u/ extracted from the MRI data of a male participant and a female participants were inserted into the head of the HATS. Two versions of each replica, with and without lips, were designed. Alternatively, as a reference, a small loudspeaker was used to simulate the mouth. The radiation patterns were measured in an anechoic room.

2.1 Experimental setup

The HATS was built using two spheres with diameters of 21 cm and 51 cm to simulate the head and the torso, as shown in Figure 1. The spheres were made of polystyrene covered with a 5 mm coating of adhesive and reinforcing mortar (Maxit multi 300) to increase the reflectivity of the surface. A tube made of unplasticized polyvinyl chloride (uPVC) with a diameter of 11 cm connected the two spheres and simulated the neck. 3D-printed vocal tract replicas could be placed inside the head to simulate the vocal tract. A sound source was attached at the glottal end of the replicas using clips. The connection was made airtight using a soft silicone seal. The sound source was a loudspeaker enclosed in a casing whose design is described in [29]. Since this sound source was not calibrated, the spectral properties of the vocal tract replicas could not be rigorously measured. However, this was not a problem since the measured data were normalized (see Sect. 2.2). Alternatively, a loudspeaker (VECO 32KC08-1A) could be placed on the surface of the sphere used to simulate the head without using a vocal tract replica. The head was filled with sand in order to absorb as much as possible the vibrations generated by the sound source, as well as the sound potentially radiated from the walls of the replica. To further reduce the potential transmission of sound through paths other than the mouth, the torso was filled with sound-absorbing material and modeling clay was used to seal the gaps between the various elements of the HATS.

The contribution of the secondary transmission path to the overall radiated sound was estimated by measuring the sound radiated when the mouth was closed with modeling clay (see Fig. 4 in the Supplemental material). The sound radiated was overall about 20 dB lower than it was when the mouth was open. Above 5 kHz, no sound distinguishable from the background noise was detected. The interference with the mouth radiation was almost completely negligible, except at some particular frequencies at which the mouth radiation had a very low amplitude. This is further discussed in Section 4.

\[ \text{Figure 1. Schematic of the head and torso simulator (figure adapted from [28]).} \]
geometries corresponding to participants 1 (male) and 2 (female) from the database were modified to integrate them into the HATS and to create two versions with and without lips. Fixations were added to attach these vocal tract replicas inside the HATS and a spherical cap was added around the mouth opening to integrate them into the head of the HATS. In the case with lips, this flange was positioned at the corner of the lips. For the case without lips, it was positioned halfway between the corner of the lips and the most external part of the lips, before the lips removal. This makes the effective length of the vocal tract about the same for replicas with and without lips [31]. The part of the lips outside of the flange was removed afterwards. The remaining gaps in front of the corner of the lips were filled manually using the sculpture function of Blender [32]. The mesh processing was done with Blender and Meshmixer. The replicas were 3D-printed with Ultimaker UM3 using polylactic acid (PLA) with 100% filling.

Figure 3 presents a schematic of the measurement setup. The HATS was placed on a turntable (LinearX LT360) that could automatically rotate the HATS to specific azimuthal positions spaced by 15°. The sound was recorded with a measurement microphone (MTG MK 250) attached to a circle arc support, allowing the microphone to be placed at seven different elevations spaced by 15°. The microphone support could be flipped down, as illustrated by the dashed lines in Figure 3, to measure positions ranging from 0° to 165° in the vertical plane (0° being above the top of the head). The microphone was connected to the Klippel Distortion Analyzer 2, which was also used to generate the source signal; this source signal was amplified by a Samson Servo 120a amplifier. The output voltage of the amplifier was measured by the Klippel Distortion Analyzer 2 before being fed to the sound source. The measurement process was controlled with a laptop using the Klippel Robotics software. The measurements were made in the anechoic room of TU Dresden.

2 We deliberately do not specify the gender hereafter to avoid any gender interpretation of the results since studying only one specimen of each gender is not sufficient to allow us to draw conclusions regarding this parameter. We interpret the differences simply as the results of two different anatomies.

### Table 1. Width and height in cm of the mouth exits of the 3D-printed vocal tract replicas.

<table>
<thead>
<tr>
<th></th>
<th>Participant 1</th>
<th>Participant 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No lips</td>
<td>Lips</td>
</tr>
<tr>
<td>Width</td>
<td>3.4</td>
<td>5.0</td>
</tr>
<tr>
<td>Height</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>Width</td>
<td>3.6</td>
<td>4.0</td>
</tr>
<tr>
<td>Height</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Width</td>
<td>1.1</td>
<td>1.0</td>
</tr>
<tr>
<td>Height</td>
<td>0.7</td>
<td>0.4</td>
</tr>
</tbody>
</table>

2.2 Data processing

The generation of the input signals and the processing of the microphone signal were carried out with the software Klippel dB-Lab RnD. The input signal was an exponential sine sweep that generated frequencies from 100 Hz to 20 kHz in 1.4 s.

A problem specific to speech directivity is that at some angular positions and frequencies, no data but background noise can be recorded. This is obviously the case for recordings of human participants, but it is also the case for the experiments presented in this study because, similarly to the vocal tracts of human participants, the vocal tract replicas substantially damp some frequencies. This is illustrated in Figure 4a, in which the normalized amplitude radiated in front of and behind the HATS is shown, as well as the smoothed background noise. One can clearly observe the effect of the filtering from the vocal tract replica, which induces peaks and troughs in the radiated sound. For this reason, the measurement was repeated four times at each position, and the background noise was recorded for each elevation position of the microphone (but not for each azimuthal position). The signal spectra were computed using a half-Hanning window, resulting in 67,200 samples (the sampling rate was 48 kHz). The spectra were averaged over the four repeated measurements. They were exported as text files for further processing using MATLAB. To reduce the amount of data, the spectra were decimated by a factor of 8 and the data beyond 20 kHz were discarded. These data are provided in the Supplemental material.

Because of the very fine frequency resolution used, the spectra had many small fluctuations (even after decimation). These fluctuations were reduced through smoothing using the MATLAB function smoothdata with window lengths of 12 and 125 samples for the radiated sounds and the background noise, respectively. More samples were used for the background noise to obtain a smoother noise threshold, which was used to exclude the noisy data for a better visualization and to evaluate the reliability of the directivity index (DI) computation (see Sect. 2.3).

To observe the amplitude variation purely related to the directivity, it is necessary to normalize the data to eliminate the overall amplitude variations related to the acoustic properties of the sound source and the vocal tract. This was done by dividing the measured amplitude by the overall maximum for each frequency. The normalized amplitude $P$ is expressed as

$$P(\omega, \theta, \varphi) = \frac{P(\omega, \theta, \varphi)}{\max[P(\omega, \theta, \varphi)]}.$$  (1)

The data were spatially interpolated to enhance the visualization and to obtain a more accurate localization of the maximum. The angular spacing was reduced from 15° to 5° using the MATLAB function griddata with cubic interpolation.

The radiation patterns were visualized using directivity balloons and directivity maps. The directivity balloons (Figs. 5, 7 and 12) show the amplitude relative to the reference maximum level (normalized to 0 dB), which is
represented by the radius of each sphere. For a better visualization, the relative level is also represented using a color scale. The directivity maps (Figs. 6, 9 and 10) show how the normalized amplitude as a function of the frequency and the angular position (polar or azimuthal) in the vertical or horizontal plane. The data below the background noise level are represented in deep blue.

2.3 Directivity index

The DI is used to quantify the directionality of a sound source, which refers to how much the maximum amplitude differs from the average amplitude. Usually, for sound sources such as loudspeakers, the maximum amplitude is located in front of the source, on the symmetry axis. Hence, the DI is computed as the ratio of the amplitude in front of the source and the average amplitude \[ \text{DI} = \frac{P_{\text{front}}}{P_{\text{avg}}} \]

However, speech radiation patterns are more complex than those of simple sound sources such as loudspeakers, and the maximum amplitude often occurs in other directions \[ \text{MDI} \]

The consequence of this is that the usual definition of the DI fails to quantify the directionality, yielding, in extreme cases, irrelevant negative values.

Therefore, in this study, we propose a more relevant definition for speech. We call this modified DI (MDI). The overall maximum amplitude is used instead of the amplitude in front of the source:

\[
\text{MDI}(\omega) = 20 \log \left( \frac{\max(P(\omega, \theta, \phi))}{\bar{P}(\omega)} \right),
\]

where \( \bar{P} \) is the average amplitude

\[
\bar{P}(\omega) = \frac{1}{\Omega} \sum_{p=0}^{N_p} \sum_{u=0}^{N_u} \Delta_\Omega \Delta_\phi \bar{P}(\omega, \theta_p, \phi_u),
\]

where \( \Delta_\Omega = (\cos(\theta_p - \Delta_\theta/2) - \cos(\theta_p + \Delta_\theta/2)) \Delta_\theta \) is the solid angle corresponding to each measurement point, and \( \Omega \) is the total solid angle covered by the measurements; the measurement points are regularly spaced over \( \theta \) and \( \phi \) with spacings of \( \Delta_\theta \) and \( \Delta_\phi \), respectively.
Along with the MDI, the elevation \( \theta_{\text{max}} \) and the azimuthal position \( \varphi_{\text{max}} \) of the maximum amplitude are extracted, providing additional useful information.

If the MDI is computed by including data under the noise threshold, it will be underestimated because the average amplitude \( \overline{P} \) will be overestimated. However, excluding the data with a poor signal-to-noise ratio will further underestimate the MDI because \( \overline{P} \) will be further underestimated. This is illustrated in Figure 4b, in which the MDI computed by excluding the data under the noise threshold is, in some frequency intervals, substantially lower than the MDI computed with all the data available. Thus, it appears to be preferable to use all the data available for the computation of the MDI.

However, if too many positions have an amplitude lower than the noise threshold, the computed MDI is not meaningful. The difference between \( P_{\text{tot}} \), which is computed using all the data, and \( P_{\text{iso}, \text{noise}} \), which is computed by excluding the data below the noise threshold, is a good indicator of the reliability of the data (see Fig. 4c). Thus, it was decided that the MDI would be considered reliable when \( P_{\text{iso}, \text{noise}} - P_{\text{tot}} < 1 \text{ dB} \). The MDI values that satisfy this criterion are highlighted with a thick orange line in Figure 4b.

One can see that this process excludes a substantial amount of data, particularly at high frequencies. This is because at high frequencies, the average amplitude is generally lower and the difference between the front and back is greater. Thus, more data are below the noise threshold, especially behind the head. To obtain a reliable MDI for a larger frequency range, only the front hemisphere was used to compute the MDI. A comparison between the MDI values computed on the full sphere and those computed on the front hemisphere showed that both MDI values follow a very similar trend, but the front hemisphere MDI is, as expected, lower. A comparison using all the data measured showed that the average difference is on the order of 3 dB.

### 2.4 Simulations

The directivity of the three vowels tested was simulated with the software VocalTractLab3D\(^3\) [34]. This software computes the 3D acoustic field inside the vocal tract using the multimodal method, which relies on the decomposition of the acoustic field into the local transverse modes. The radiated field is computed by integrating the particle velocity distribution on the mouth exit using the Rayleigh-Sommerfeld integral. The mouth exit is set in an infinite boundary condition implemented in VocalTractLab3D prevents the simulation of the radiation on a full sphere.

### 3 Results

This section describes the observations made using the measurements and the simulations. As a substantial amount of data has been collected, only the most relevant data are shown to illustrate the observations described. For further data visualization, the reader is referred to the Supplemental material. The directivity maps, MDI values and directivity balloons can be plotted with the MATLAB script provided. The radiation of all the configurations with a torso, with and without lips, was measured on a full sphere, i.e., varying both the polar and azimuthal angles. However, it was considered sufficient to measure only two configurations without lips and without a torso on a full sphere (the vowel /a/ of participants 1 and 2). Nevertheless, this configuration was measured in the horizontal plane only (\( \varphi = 90 \)) for the other phonemes (/i/ and /u/).

In almost all the directivity maps, there is a frequency interval around 6 kHz with a lower amplitude than the background noise for all the angular positions. This overall poor signal-to-noise ratio can be attributed to a strong damping of the vocal tract replicas in this frequency range.

#### 3.1 Effect of torso and lips

The data measured with the torso and the lips have been averaged for both participants and all the phonemes in order to compare them with measurements of human participants published in other studies. The balloons corresponding to the third octave bands are presented in Figure 5. The directivity patterns become more directional and complex as the frequency increases. Above 1 kHz several lobes are visible; their direction varies with the frequency.

The effect of the torso and the lips is illustrated with the data measured with and without a torso and with and without lips for the vowel /a/ of participant 2. This effect is shown with directivity maps in Figure 6, directivity balloons at 1.5 kHz in Figure 7 and a more synthetic view with the MDI and the angular position of the maximum in Figure 8.

When the torso is present, it generates an interference pattern that contains repeated maxima and minima. This is referred to hereafter as the torso pattern (TP). This pattern is visible in the vertical plane (Fig. 6a vs. Fig. 6b), as well as in the horizontal plane (Fig. 6d vs. Fig. 6e). The torso affects the location of the maximum amplitude in the vertical plane, which tends to be located at the maxima of the TP, as can be seen in Figure 8b. In the horizontal plane (Fig. 8c), up to about 4 kHz, the azimuthal position of the maximum varies greatly when the torso is present. Above 4 kHz, the cases with and without

---

\(^3\) https://vocaltractlab.de/index.php?page=vocaltractlab-download
the torso have fewer variations and show a similar trend. On the directivity balloons, one can see that the torso can induce a minimum in front of the HATS (see Fig. 7a vs. Fig. 7b). This is the case at 1.5 kHz, as illustrated in Figure 7, but it is also the case at other frequencies such as, for example, those above 5 kHz (see Fig. 6c). One can also note that in all the configurations, even without the torso, a minimum is present at an elevation of about 120°.

The presence of the torso increases the front-back difference. This is visible in the balloons of Figure 7, but it is also visible in the directivity maps (Fig. 6a vs. Fig. 6b); one can see that this happens at all the measured frequencies. This results in a globally higher MDI (on average 1 dB) when the torso is present (see Fig. 8a). However, this difference is not uniformly distributed over the frequencies; there is a substantial difference of up to about 2 dB in the range 0.5–2 kHz, there are fewer differences in the range 2–4 kHz and there are again more differences above 4 kHz. For participant 1, the same tendency is observed up to about 10 kHz, after which the configurations with and without the torso have equivalent MDI values (Fig. 2 in the Supplemental material).

When the lips are present, the TP appears to be more pronounced: the difference between the maxima and the minima is greater. This is visible in both the vertical plane (Fig. 6b vs. Fig. 6c) and the horizontal plane (Fig. 6e vs. Fig. 6f), as well as on the balloons (Fig. 7b vs. Fig. 7c). The lips also increase the MDI, which is on average 1 dB higher than it is without lips and 2 dB higher than it is without lips and without a torso (see Fig. 8a). However, this increase occurs mainly at high frequencies above about 4 kHz, except in the range 9–12 kHz, in which the cases

Figure 5. Third octave band directivity balloons for the phonemes /a/, /i/ and /u/ of both participants in the configuration with a torso and lips.
with and without lips have a similar MDI. One can also see differences in the location of the maximum amplitude, which tends to be located at a higher position (smaller elevation angle) with lips in the range 6–14 kHz (see Fig. 8b). For participant 1 (Fig. 2 in the Supplemental material), this happens in the range 4–9 kHz, after which the trend is reversed; the configuration without lips has a maximum amplitude located at a higher position.

### 3.2 Effect of the vowels

The effects of vowels are illustrated with directivity maps of simulations and measurements in the horizontal plane in Figures 9 and 10. A more synthetic view is shown for participant 1 with the MDI and the angular position of the maximum in Figure 11. Figure 12 illustrates more specifically an aspect of the HOME.

Since VocalTractLab3D cannot simulate the effect of the lips, head and torso, the simulations were compared to the closest experimental case, which is the head without lips and without a torso. As one can see in Figure 9, there is an overall very good qualitative agreement between the simulations and experiments. Above 4 kHz, the detailed shape of the directivity patterns is successfully reproduced by the simulations up to about 12 kHz. See, as an example, the vertical streaks of the phoneme /a/ around 7 kHz and 8 kHz, which are visible for both the simulation (Fig. 9a) and the experiment (Fig. 9b). For the phoneme /i/, the lobe becoming narrower as the frequency increases and eventually dividing into two lobes is also successfully reproduced (Figs. 9c and 9d). However, these features occur at different frequencies in the simulations and the experiments. Thus, a quantitative comparison would result in substantial differences. However, substantial differences exist in the range 0–4 kHz. They consist mainly of localized minima of the experimental data, which contrast with the rather uniform patterns of the simulations. They are very faint for /a/ (Fig. 9b) but much more pronounced for /i/ and /u/ (Figs. 9d and 9f).
In Figures 9 and 10, there are substantial variations in the directivity patterns within small frequency intervals (on the order of 100 Hz). This is particularly visible for the vowel /a/ of participant 1 (Fig. 10a); it appears as vertical streaks (e.g., around 4 kHz, 6 kHz and 8 kHz) and as generally complex radiation pattern variations at higher frequencies. This effect is not observed when the mouth exit is simulated by a loudspeaker (Fig. 10g). Based on previous works [14, 21, 22] and the present results, this phenomenon is probably mainly caused by the HOME. This phenomenon consists of the propagation of higher-order transverse modes inside the vocal tract, which causes a non-uniform field distribution over the mouth exit and significantly affects the directivity of the radiated sound. It is more evident in the horizontal plane, but it is also observed in the vertical plane; see, as an example, Figures 6a, 6b and 6c in the range 11–12 kHz. The consequences of the HOME are also visible in the azimuthal position of the maximum as deviations from 0° and abrupt changes in the azimuth above 4 kHz (see Figs. 8c and 11c).

The HOME is illustrated in more detail in Figure 12 with directivity balloons created using measured and simulated data. The selected frequencies correspond to the HOME streak visible near 5 kHz in the directivity maps of the simulated (Fig. 9a) and measured (Fig. 9b) data for the phoneme /a/ of participant 2 without lips and without a torso. In this example, both the simulation (Fig. 12c) and the experiment (Fig. 12d) show an asymmetric radiation pattern divided into two lobes, whereas the
piston radiation model would show only one symmetric lobe in the absence of the HOME (see Fig. 1a vs. Fig. 1b in the Supplemental material). The simulated amplitude of the acoustic particle velocity on the mouth exit surface is lower in the center (Fig. 12a). Its phase (Fig. 12b) is almost symmetrically divided into two areas with nearly opposite phases (a difference of about \( \pi \)). Note that this is just one illustration of the HOME and that in other cases, different field distributions can be observed.

The HOME is more visible for /a/ (Figs. 10a and 10b), less visible for /i/ (Figs. 10c and 10d) and almost absent for /u/ (Figs. 10e and 10f). For /a/, it is visible starting at a lower frequency (about 4 kHz) for participant 1 (Fig. 10a) compared to participant 2 (about 5 kHz; see Fig. 10b). For /i/, it is almost absent for participant 2 (Fig. 10d), but it is observable for participant 1 (Fig. 10c). For /u/, it is totally absent for participant 1 (Fig. 10e) and visible only as narrow vertical streaks for participant 2 (Fig. 10f). For /i/, the HOME affects the directivity in a broader frequency range. Within 10–15 kHz for the simulation and 12–15 kHz for the experiment, the directivity is reduced in comparison with the piston radiation model (Figs. 9c and 9d; see also Figs. 1d, 1e, and 1f in the Supplemental material). This effect is more difficult to identify in the measurements with lips and a torso; however, there is a decrease in the MDI in the range 10–12 kHz (see Fig. 11a), which could be related to this effect. Note that there is also a decrease in the MDI, to a lesser degree, for /a/ in the same frequency range.

When the lips are present, the HOME is slightly less visible. See, in particular, the streak around 7 kHz in Figure 6e, which is almost invisible when the lips are present in Figure 6f.

In Figure 10, one can see that the TP varies for different vowels. For /a/, the structure of repeated maxima and minima is less visible toward the high frequencies (starting from about 14 kHz). It is more visible for /i/ and even more clearly observable for /u/. One can also see differences between the participants: for example, for /i/, there is a deep minimum around 8 kHz for participant 2 (Fig. 10d), but for participant 1 (Fig. 10c) this minimum is less pronounced and occurs at lower frequencies.

The vowels /a/ and /i/ are substantially more directional than /u/ (MDI up to 4 dB higher), particularly in the range 4–10 kHz for participant 1 (see Fig. 11a). A similar but less pronounced trend is observed for participant 2 (see Fig. 3 in the Supplemental material). At high frequencies, starting from about 14 kHz, /a/ is substantially more directional than /i/ (MDI up to 4 dB higher), but there is no substantial difference below this frequency for participant 1. For participant 2, /a/ is slightly more directional than /i/ (about 0.5 dB) in the range 4–11 kHz (see Fig. 3 in the Supplemental material). One can also notice some peaks of the MDI for /u/ between 1 kHz and 3 kHz. They are at the edges of the low signal-to-noise-ratio frequency ranges; see the deep blue vertical streaks in Figure 10e, which indicate that a large amount of data lay under the noise threshold. Thus, the relevance of these peaks can be questioned.

The location of the maximum in the vertical plane (Fig. 11b) is very similar for all the phonemes. However, there are a few substantial differences, for example, in the range 7–9 kHz, in which there is a lower position for /u/. In the horizontal plane (Fig. 11c), the location of the maximum varies greatly without any phoneme-specific trend up to about 3 kHz. Above 3 kHz, the variations of /u/ are more dispersed and /a/ and /i/ have different types of deviations from the central position of 0°.

Comparing the cases with and without lips for each vowel, the MDI difference between the configurations with and without lips becomes visible at different frequencies depending on the phoneme and the participant. For /a/,
4 Discussion

The third-octave averaged directivity patterns can be compared to the measurements of Leishman et al. [7] on human participants (see, in particular, Figs. 14, 18 and 20 of [7]). Similar directivity patterns are observed up to 250 Hz. At higher frequencies, qualitatively similar patterns can be observed in both experiments: the directivity patterns divide into several lobes in both cases, but they have different orientations and widths, and/or similar patterns occur in different frequency bands. These differences can be attributed to the different dimensions of the torso of the HATS and the torso of the human participant, the limitations of the geometrical approximation used for the HATS and the restriction of our study to three vowels, whereas the study of Leishman et al. was done using running speech. Nevertheless, the HATS reproduces the main features of human speech directivity.

The TP observed is qualitatively similar to that reported by Brandner et al. [15], and it corresponds to the observations of human participants [16]. Thus, a simplified HATS can reproduce the main features of the TP. Our results complement the previous results by showing the TP in the vertical plane and with directivity balloons, as well as by investigating how it is coupled with other parameters such as the lips, the mouth area and the HOME. It appears to be more pronounced when the lips are present. The reason for this is not well understood, but it could be related to the stronger directionality induced by the lips, which could reduce the importance of the diffraction due to the head with respect to the diffraction and reflections due to the torso. The variations of the TP with the phone are related to both changes in the mouth dimensions and the HOME, which is superimposed on the TP and can mask it partially (completely starting at about 14 kHz in the case of /a/). The case of the vowel /i/ shows that the HOME can also modify the TP on wide frequency intervals (3 to 5 kHz wide). This is probably due to a change in the particle velocity distribution at the mouth exit, which exists on a wide frequency interval because of the specificities of the vocal tract shape of /i/. Finally, it is worth mentioning that if the sphere used to represent the torso had a different size, this would lead to a different TP. This was suggested in the interpretation of the differences between radiation patterns measured from different participants [15].

The differences in the MDI observed between the cases with and without lips can be in part attributed to the differences in mouth dimensions. In particular, for /a/, the mouth width without lips is substantially smaller (see Tab. 1), which further increases this difference. However, the configurations without substantial changes in the mouth dimensions (/i/ of participant 2 and /u/) also show an increase in the directivity in the case with lips. This is in line with the findings of Yoshinaga et al. [13]. The frequency of 5 kHz proposed by Yoshinaga et al. as the lower frequency limit of this effect also corresponds globally to the order of magnitude observed in our results. We observed that the lips affect the directivity in various ranges depending on the phone (from as low as 2 kHz for /a/). The general trend is that the larger the mouth opening, the lower the starting frequency at which the influence of the lips is visible. One can also note that there can be frequency intervals at higher frequencies in which the directionality is not affected (mostly in the range 9–12 kHz). However, to better discriminate between the effects of the mouth dimensions and the presence of the lips, one should make sure that the mouth opening dimensions are similar in both cases. The lips also tend to shift the maximum amplitude toward the top of the head in the range 3–14 kHz. The reason for this is not clearly understood and a proper model of the effect of the lips would be necessary to study it.

The relationship between the mouth dimensions (related to the different phonemes) and the directivity is globally verified. One can see in Table 1 that the mouth dimensions of /a/ are the largest and those of /u/ are the smallest (for the case with lips). Thus, the larger the mouth dimensions, the stronger the directivity. This is easily explained by the simple piston model [9, 11], and it is in agreement with observations of human participants [33]. However, these differences in directivity are observed starting at about 4 kHz, which is slightly higher than the starting frequency for human participants, for whom these differences were observed in the octave bands 2 kHz and 4 kHz [16, 17] and starting at about 1 kHz [35] (with a different definition of the DI). The substantial variations in the azimuth of the maximum up to 3 kHz are probably due to the low directionality in this range: small fluctuations of the amplitude impact the azimuthal position of the maximum more, while the main variation of the amplitude is related to the elevation.

The good qualitative agreement between the simulations and the experiments shows that the radiation model implemented in VocalTractLab3D successfully reproduces the effect of the vocal tract and the mouth opening size, and it approximates the diffraction caused by the head well, even though, obviously, it cannot predict the acoustic field behind the head. The fact that this model cannot be compared quantitatively to experiments can be attributed to
the small remaining geometrical differences between the vocal tract replica and the simulated geometry. In fact, the HOME is very sensitive to small geometrical differences. It was observed by Motoki that small perturbations of a simplified vocal tract geometry induced substantial differences in the simulated transfer function above 4 kHz [36]. However, some substantial differences were observed in the range 0–4 kHz. The origin of these differences is not very well understood, but they might be the consequence of reflections that were insufficiently damped, as an example, on the support of the HATS (even though some damping material was placed on it). The minimum at about 120° observed in Figure 7 supports this hypothesis. However, the minima observed in Figures 9b, 9d and 9f vary with the phoneme, so the cause of these minima is probably more complex than unwanted reflections. This phoneme dependency could mean that some sound was transmitted through paths other than the mouth exit, e.g., through the vocal tract replica walls or through the walls of the sound source and the HATS. These secondary transmissions may mainly impact the radiated sound in frequency ranges with particularly low radiated amplitudes. This could also explain why /a/ is much less affected than /i/ and /u/. This hypothesis is supported by the comparison between the mouth being open and the mouth being closed with modeling clay (see Fig. 4 in the Supplemental material). It was observed at some specific frequencies that the sound radiated by the secondary paths could be up to 10 dB greater than the sound radiated by the mouth at some specific frequencies and angular positions. By comparison with the directivity map, one can see that these frequencies correspond to unexpectedly complex patterns in the range 1–4 kHz. However, when the data below the noise threshold are removed, most of these patterns are no longer visible. This indicates that the main problem is the very low amplitude of the radiated sound in some frequency intervals due to the filtering of the vocal tract replica. The sound radiated by the secondary path still has overall a very low amplitude, as illustrated in Figure 1b of the Supplemental material. Thus, this phenomenon should be investigated with care in future studies. This also raises the question of the existence of secondary transmission paths for human speakers. Radiation from the chest, the larynx area, the nose or the cheeks may induce similar interference patterns for frequency ranges at which the amplitude from the most direct path is particularly low.

The HOME was clearly identified and further documented. Our observations are in good agreement with previous observations [16, 23] and theoretical studies [14, 22]. In particular, the fact that the HOME is the most visible for /a/ and the least visible for /u/ is confirmed experimentally. The /i/ data show that the HOME can affect the directivity continuously in a relatively broad frequency range (around 5 kHz) for some specific geometries. The simulations performed with VocalTractLab3D allowed us to explore in further detail the relationship between the particle velocity distribution at the mouth exit and the directivity patterns induced by the HOME. It is confirmed that the plane piston model is not valid in these particular cases since both the amplitude and the phase substantially vary over the mouth exit. Furthermore, a division of the mouth exit into two areas with opposite phases confirms the intuitively expected origin of the two-lobes pattern. It is also noteworthy that the HOME can substantially reduce the directionality, in particular for /i/. Finally, it was found that the presence of the lips tends to make the HOME less pronounced. The reason for this is not clearly understood, and a proper model of the effect of the lips would be required to study it.

Some substantial differences have been found between the two participants studied. However, the origin of these differences is probably manifold: it could be related to the gender (participant 1 is male and participant 2 is female), inter-individual differences, the articulation characteristics at the moment of the MRI recording (participant 2 had a more pronounced articulation) or the geometry extraction process performed on the MRI data. Further studies would be required to estimate to what extent each of these parameters plays a role. Nevertheless, some differences can be related to geometrical differences. The most straightforward is the difference in the mouth opening dimensions, which can be directly related to differences in directionality. For example, one can see in Table 1 that the mouth width for /i/ is smaller for participant 2 and, as expected, it is globally less directional. The participant-specific mouth dimensions also explain why more differences in the MDI are observed between /i/ and /a/ for participant 2. Participant 2 has a more pronounced articulation for /a/: the mouth height is twice as large as that of participant 1, while the width is the same (see Tab. 1). In addition to the relationship between directionality and the mouth dimensions, the associated greater directionality can be related to the observation of Brandner et al. [15] that a lower position of the jaw induces a stronger directionality and more downward radiation at high frequencies. Concerning the HOME, its appearance at lower frequencies for the vowel /a/ of participant 1 can be related to his larger interdental spaces (see Fig. 5 of the Supplemental material). The larger interdental spaces lower the cutoff frequency of the higher-order modes, which causes the HOME to appear at lower frequencies.

5 Conclusion

A good qualitative agreement was observed between the directivity patterns of the HATS, the simulations and the measurements of human participants from Leishman et al. [7]. However, some low-frequency differences between the simulations and the HATS need to be better understood. The origin of the TP was confirmed and it was described in further detail. It was found that the TP was enhanced by the presence of the lips and can be more or less masked by the HOME. It was confirmed that larger mouth dimensions induce a stronger directionality. The increase in the directionality induced by the lips was also confirmed, and it was shown that the frequency range of this effect is related to the phoneme being pronounced and likely to the mouth opening dimensions as well. The HOME was...
experimentally confirmed and documented in further detail. This phenomenon, described earlier as a substantial variation of the directivity pattern within small frequency intervals, was also shown to affect the directivity more continuously on larger intervals (on the order of 5 kHz). The simulations illustrated in further detail how the HOME affects the directivity. In particular, the division of the mouth exit surface into two areas with nearly opposite phases was illustrated. We observed that the HOME can be attenuated by the lips. Finally, it was observed that differences in the vocal tract geometry of a given phoneme can impact the directionality and the HOME.

Data availability statement

The supplementary material can be downloaded at: https://vocaltractlab.de/index.php?page=birkholz-supplements [37].

Conflict of interest

Author declared no conflict of interests.

Acknowledgments

This study was supported by the German Research Foundation (DFG) with grant no. BI 1639/7-1.

References


Cite this article as: Blandin R. Geng J. & Birkholz P. 2023. Investigation of the influence of the torso, lips and vocal tract configuration on speech directivity using measurements from a custom head and torso simulator. Acta Acustica, 7, 39.