Pilot study on the influence of spatial resolution of human voice directivity on speech perception

Aurian Quélennec, Paul Luizard

Acta Acustica 2022, **6**, 10

https://doi.org/10.1051/aacus/2022006

Available online at:
https://acta-acustica.edpsciences.org

SHORT COMMUNICATION OPEN ACCESS

Pilot study on the influence of spatial resolution of human voice directivity on speech perception

Aurian Quélennec[1] and Paul Luizard[1,2,*]

[1] Audio Communication Group, Technische Universität Berlin, Einsteinufer 17c, 10587 Berlin, Germany
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France

Received 16 October 2020, Accepted 14 February 2022

Abstract – A perceptual threshold related to spatial resolution of the human voice directivity was determined through a listening test of similarity (MUSHRA). Directivity data of an artificial talking head measured at high spatial resolution (spherical harmonics order 35) was the input of a room acoustics simulation software (RAVEN) to build sound stimuli in various room acoustic conditions and source–receiver arrangements, with different voices. Results showed that, at spherical harmonics order 8 and above, the voice signal was not anymore perceived as significantly different from the greatest resolution. An analytical model was proposed and showed good agreement with the listening test results.

Keywords: Voice spatial resolution, Room acoustic simulation, Voice perception, Perceptual threshold

1 Introduction

Most previous research on voice directivity investigated technical aspects through measurements of human beings [1, 2], either for random speech [3] or specific sung vowels [4], and also artificial heads [5, 6] that present the advantage of repeatability. The perception of the human voice directivity was so far seldom studied, although most human beings are able to hear differences when a speaker is facing them or turned in the opposite direction. Previous research [7] on perception of voice directivity emphasised the influence of high frequency, in the octave bands 8 and 16 kHz, especially on speech localisation and intelligibility in rooms. Regarding the concept of immersion in virtual reality, the degree of improvement yielded by implementing voice directivity in the context of interaction between a virtual agent and a user was evaluated [8]. Results did not indicate significant effects regarding directivity on the perceived social presence and on the realism of the virtual voice. However, the authors mentioned that these results were mainly due to other aspects of the experiment and indicated possible design options to highlight the effects of voice directivity in further research.

A previous study [9] has found a threshold for human voice directivity using sound stimuli of various spatial resolution values. The threshold was expressed in terms of spherical harmonics (SH) order that was used to synthesize the sound stimuli. It was found to be order 5. It should be noted that this experiment was conducted using sound stimuli generated using a custom voice directivity, not measured data, and in a single arrangement where the receiver faces the speaker. The present study intends to investigate the threshold for spatial resolution of human voice directivity based on measured data and in various arrangements of the simulated source and receiver, i.e. several speaker head orientations and voice types, in anechoic and reverberant rooms.

In this study, three technical assumptions are made on the directivity of an artificial head, on the order reduction strategy, and on the head related transfer function (HRTF) filters: (1) Directivity of artificial speakers, such as the *Head Acoustics HMS-II,* presents similar characteristics as human speakers according to a previous work [6]. (2) An order reduction by simple truncation of the coefficient of the spherical harmonic transform is an efficient method [9] for sound stimuli synthesis. (3) The simulations conducted with the RAVEN software include HRTF filters to obtain an auralization in binaural format. The HRTF used in the present study are measurements of the FABIAN dummy head developed at TU-Berlin [10].

The experimental hypothesis tested with this perceptual test is that variations of the spatial resolution of binaural renderings of the human voice can be perceived up to a threshold above which no difference can be heard anymore between higher resolutions. This threshold is intended to be quantified with respect to various external parameters.

It should be noted that this study involves static configurations of a speaker and a listener, where the latter is not able to move the head. The spatial auditory cues involved

*Corresponding author: paul.luizard@lilo.org

are the directivity patterns of the sound sources and the HRTF of the listener. The sound stimuli constructed using those filters allow for various degrees of externalisation (feeling that the sound surrounds the listener, not only located inside the head) of the sound scenes for the different participants of the listening test. The ability to move the head represents a temporal succession of speaker–listener configurations, i.e. various transfer functions to construct the stimuli. Although this solution would have been slightly beneficial to the naturalness of the sound stimuli, it requires a head orientation tracker, which would have made it impossible to run the listening test online.

2 Method

In order to determine the spatial resolution order where the listener starts hearing a difference between the reference version at full spatial resolution and a degraded one, a listening test was implemented. Two different methods, recommended by the International Telecomunication Union, seemed to be efficient to compare the audio files of different quality in this study: the ITU-R BS.1534 (MUSHRA) [11] and the ITU-R BS.1116 [12]. The MUSHRA framework was selected because it allows for comparing a higher number of sound stimuli within the same duration of test. This method consists in comparing all sound stimuli with different degrees of degradation of spatial resolution, presented on the same panel, to the reference. In the present study, the listeners had to rate the similarity to the reference on a scale from *bad* to *excellent* with five different levels. Among the sound stimuli, one of them is the hidden reference, which should be rated close to 100%. In addition, an anchor (low-pass filtered at 3.5 kHz version of the reference) stands for a clearly degraded stimulus that should obtain a bad rating. The reference and the anchor were used to exclude outlier listeners according to the criteria described at the end of Section 2.

2.1 Sound stimuli generation

To produce the sound stimuli, directivity measurements of the artificial speaker *Head Acoustics HMS-II* at high resolution (2702 points on a Lebedev full spherical grid) was used [13]. These measurements resulted in data of 35th order in the spherical harmonics domain, according to the equation $N_m \geq (N_{\mathrm{SH}} + 1)^2$, where N_m is the number of measurement points and N_{SH} is the maximal corresponding SH order [14]. Using the spatial decomposition on the orthogonal base of spherical harmonics, data processing was performed by means of simple truncation of the coefficient of the spherical harmonic transform to obtain directivity patterns at lower order, i.e. with lower spatial resolution. The signals were then filtered into third-octave bands so that they could be used in conjunction with the boundary absorption and diffusion conditions in the RAVEN software. As a result, *OpenDAFF* files with spatial resolution at every SH order from 1 to 35 were generated. An overview of these directivity patterns is presented in Figure 1.

Those files were used in RAVEN to define the sound source directivity in order to compute binaural room impulse responses (BRIR) at the receiver positions. These BRIRs were later convolved with anechoic speech recordings from a male and a female speaker from the Harvard Word List.[1] The simulated rooms used for the experiment were a basic parallelepipedic meeting room with dimensions 5 m × 8 m × 4 m. None of the opposite faces were strictly parallel, a small angle of 1° between them was introduced to avoid flutter echoes. The uniform material in this simulated room was chosen to achieve a reverberation time of 0.7 s with a scattering coefficient of 30% across the whole frequency spectrum. The second simulated room was an anechoic chamber of identical dimensions, covered with a fully absorbent material. The arrangement of the speaker and listener presented no symmetry regarding the room. In addition to room acoustics, the head orientation of the speaker was also a variable parameter. Three different head orientations were simulated in the meeting room: the speaker and the listener facing each other (0°); the speaker making a 90° angle with the listener; the speaker making a 180° angle with the listener. In the anechoic chamber, the speaker was only facing the listener. The eight combinations of variable parameters (voice type, room acoustics, head orientation) for the sound stimuli are summarised in Table 1. Comparing the frequency spectra of stimuli of different SH orders in identical scenarios generally showed that higher SH orders yielded more energy at higher frequency than for lower SH orders, as illustrated in Figure 2 which presents the spectra of binaural room impulse responses (left side). This phenomenon can be understood by considering the directivity patterns in Figure 1 under a certain fixed angle, where the amount of received sound energy at a given frequency varies with the SH order. This occurs at each frequency point, yielding various spectra related to each SH order truncation.

2.2 Online MUSHRA test

To reach as many participants as possible, it was decided to make the test available online rather than keeping it in the lab. The user interface was recently developed by the International Audio Laboratories Erlangen and is compliant to ITU-R BS.1534 recommendations [15]. In order to keep the test duration under 30 min, the number of sound stimuli was limited to eight, plus the anchor and the reference. Therefore each trial, corresponding to a given scenario (i.e. combination of voice type, room type, and head orientation), contained 10 sounds to be rated. The selected SH orders were 1, 2, 3, 4, 5, 6, 7, 10, and 35 because the pre-tests showed that this selection provided the smoothest progression of ratings for subjective similarity to the reference (order 35). The short questionnaire administered at the end of the listening test revealed that the listeners were mostly males (25 participants out of 27) with on average 34 years of age, 23 participants reported playing a musical instrument, and 22 participants had a

[1] https://odeon.dk/downloads/anechoic-recordings/

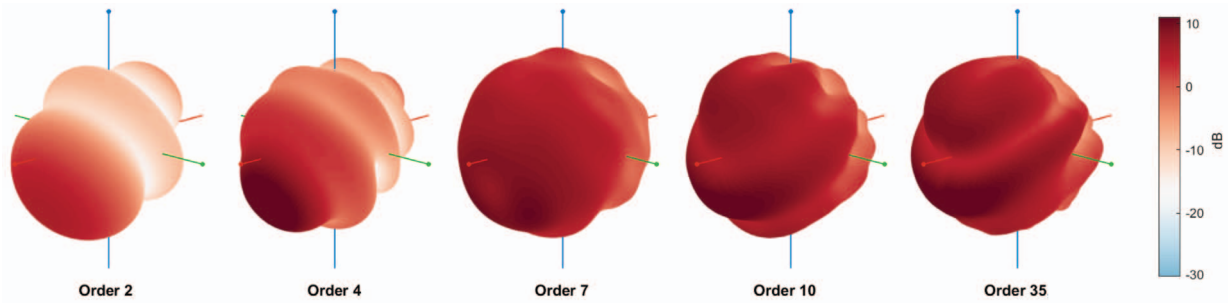

Figure 1. 3D view of directivity patterns from measurements of the artificial speaker at 3 kHz for truncation at various spherical harmonics orders.

Table 1. Spherical harmonics order at which the participants rated the stimuli significantly different than the reference (SH order 35). A to C: Male voice, meeting room, head orientation of the speaker 0°, 90°, 180°; D to F: Female voice in same conditions; G and H: Anechoic room, speaker facing listener, male and female voices resp.

Condition	A	B	C	D	E	F	G	H
SH order	7	10	6	10	10	10	7	7

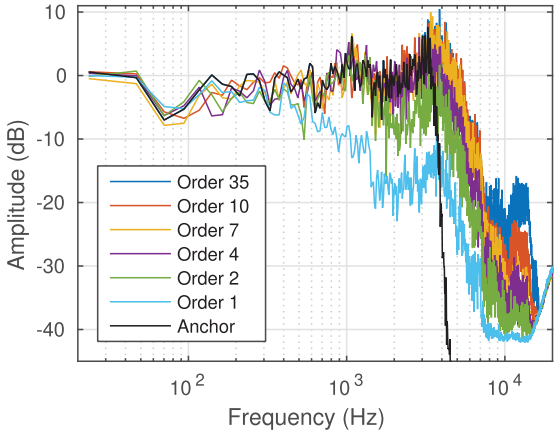

Figure 2. Spectra of the left binaural impulse responses in the meeting room when speaker and listener face each other, for various SH orders of the speaker signal and the anchor (low-pass filtered version of the signal at order 35) of the test.

professional activity related to audio or sound. Regarding previous listening test experience, 3 participants answered *not at all*, 11 *moderately*, 8 *much*, and 5 *very much*. They were recruited through mailing lists dedicated to audio.

As proposed in previous research [12], the obtained data was analysed by means of standard statistical tools such as average values, standard deviations, and confidence intervals (CI) at 95%. The latter allows for directly identifying whether two series are significantly different, e.g. to discriminate the average ratings between two SH orders. During the post-selection process of the listeners, i.e. to identify the outliers, the mean ratings of each listener for the anchor and the hidden reference were used. The reliability of responses was estimated in a classical way for a MUSHRA test [11], meaning that, for 15% of responses, the anchor must not have obtained a good score (<0.9) and the hidden reference must have been well rated (>0.9). Otherwise the participant was considered an outlier and the responses were discarded.

3 Results

Among the 27 participants, four outliers were identified, resulting in 23 remaining participants. The first general observation was that the average ratings for similarity to the reference increased with the SH order in almost all scenarios.

3.1 Threshold to the reference

The confidence intervals of different series of ratings allowed for detecting significant differences between these series which occur when two confidence intervals do not overlap. The task of the listening test being to rate the similarity level of each stimulus to the reference sound, it appeared interesting to first focus on the SH order at which a significant difference with the reference could be observed. Table 1 shows that these threshold values, depending on the scenario, range from SH order 6 to 10.

It is noted that SH order 1 provided the lowest ratings, always significantly different than higher orders. The rating values under 20% relate to a *low similarity* to the reference. In addition, above order 3, the successive confidence intervals overlap in almost all scenarios, the only exception being for the female voice in the meeting room at 0° and 90° angles. Confidence intervals of the reference were smaller than most others, but regularly overlap with order 10 in the meeting room. In the anechoic chamber, in contrast, order 10 and the reference were always overlapping.

As a general result, the mean threshold of perceptual similarity to the reference at SH order 35 is SH order 8.4.

This is a slightly higher value than the results of a previous study [9] where the average value was SH order 3 to 4 for speech and 4 to 7 for noise stimuli. This discrepancy can be due to the process of generating the stimuli that used different basis material (directivity measurements vs. custom model of directivity), auralisation algorithms (RAVEN software vs. VST plugins), and HRTF sets.

3.2 Effect of gender, head orientation, and room acoustics

This experiment had three independent variables: (1) voice type, with the parameter gender, which had two levels: male and female; (2) head orientation, with the parameter angle, which had three levels: 0°, 90°, and 180°; (3) reverberation, with the parameter presence, which had two levels: absent and present. The rated similarity to the reference was the dependent variable.

Figure 3A shows that no significant difference could be observed between the male and female voices in the meeting room when the speaker and the listener are facing each other. However in the anechoic chamber a strong trend was observed for the female voice whose ratings were always above those of the male voice, although no significant difference could be observed. In addition, the threshold of similarity to the reference is identical for both voice types.

The effect of the angle between the speaker and the receiver was investigated for both voice types. For both male and female voices, it was clear that the ratings for a 180° angle are above the others. This could be partly explained by the fact that most of the energy at high frequency is located in front of the speaker. As underlined by previous research [7], the energy at high frequency plays a great role in perception of localization of voice, its quality and intelligibility. In the case of a 180° angle where high frequencies are attenuated, the perceived difference between the reference and the stimuli was smaller at a given SH order than in the case of 0° where the high frequencies are more present.

Finally the type of room, i.e. the presence of reverberation, had a noticeable effect on the perception of similarity. The differences seemed to be less audible in the anechoic room than in the meeting room. This might be due to the lack of early reflections in the anechoic room where only the direct sound propagates. In contrast, in a reverberant room, the listener can take benefit of the early reflections, e.g. by using binaural cues [16], to perceive more subtle changes in the voice of the speaker.

In the short questionnaire, the participants were asked to indicate the extent to which they used each sound feature *Spatialisation*, *Loudness*, and *Spectral balance* to perform the similarity task. Figure 4 shows that *Spectral balance* was the most used feature. About 70% of the participants reported to use *Spatialisation* moderately to very much, while *Loudness* was less used. These results could be understood as two ways of discriminating the proposed sounds, either directly by using the spatially heard differences, and indirectly with the spectral variations which are due to the change of directivity depending on SH orders. It appears that indirect cues are mostly used here.

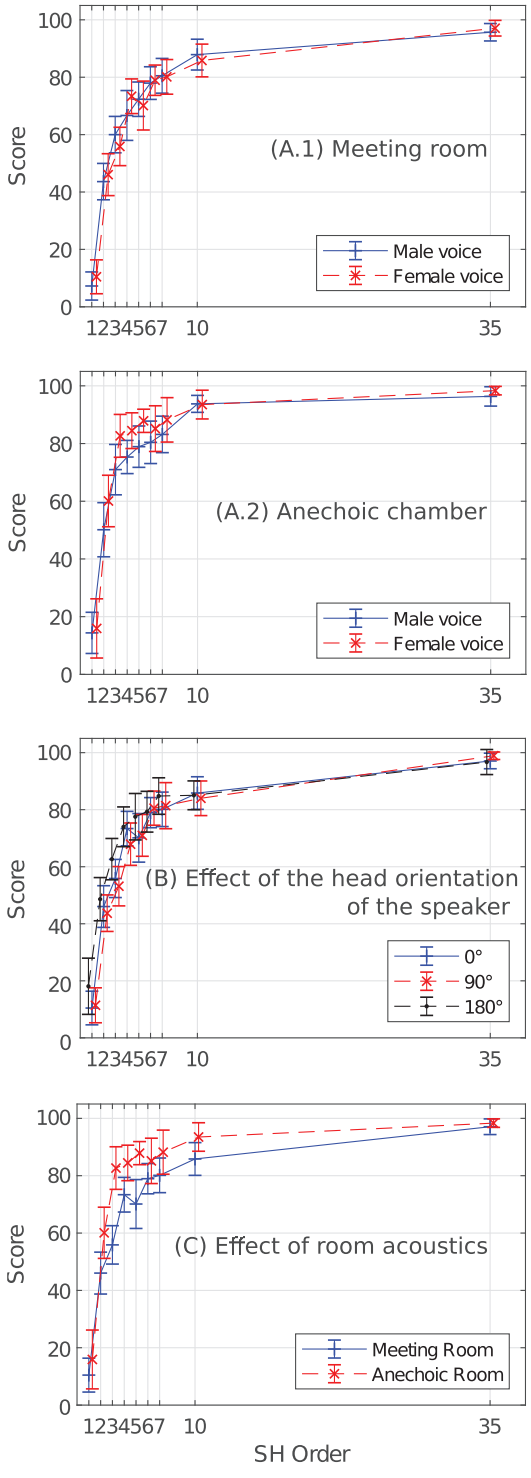

Figure 3. Average and CI 95% of the ratings for the male and female voices in (A.1) slightly reverberant and (A.2) anechoic rooms when the speaker faces the listener, (B) different head orientations, and (C) in different rooms for the female voice.

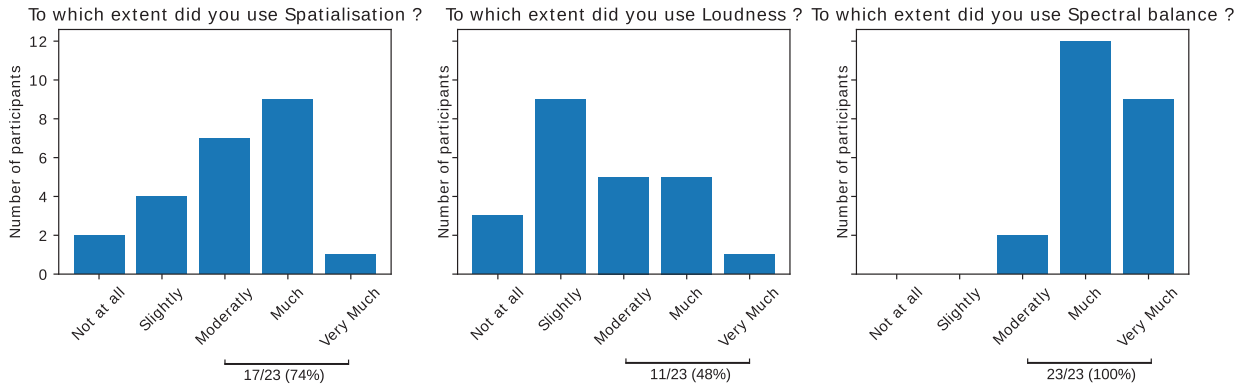

Figure 4. Number of participants who reported to use specific sound features in the similarity task.

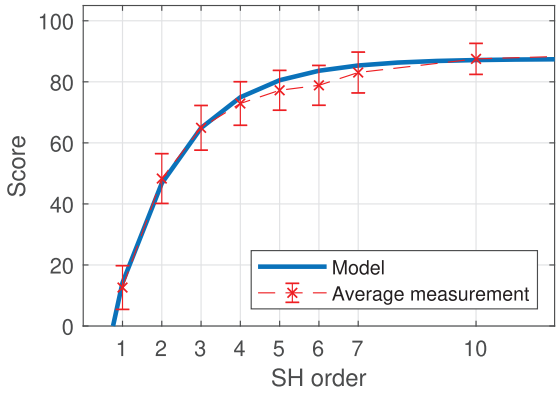

Figure 5. Average measurements with CI 95% and the logarithmic model curve.

3.3 Model of similarity

The variation of similarity ratings of the stimuli to the reference sound, depending on the SH order, is not linear. This variation can be modeled by:

$$y = m\left(1 - \exp\left(-\frac{x - 0.7}{\tau}\right)\right), \tag{1}$$

where y is the rating of similarity, x is the SH order, m is the maximal rating value, and τ is the rate of variation. Here $m = m_{35} - 9.5 = 87.5$ where $m_{35} = 97.0$ is the average rating at SH order 35, and $\tau = 1.7$. The offset value 0.7 relates to the minimal SH order, i.e. 1, not 0.

The accuracy of this model can be quantified by calculating its root mean square (RMS) deviation to the average results of the listening test. Both model and test outputs appear in Figure 5, where the model remains within the confidence intervals up to SH order 10. The global RMS deviation value of 3.9% is rather small, which confirms that this simple model provides a high level of accuracy. This shows that the perceived degradation of speech spatial rendering follows a logarithmic relationship with SH orders. In addition, the model can be used to estimate the similarity of a reduced spatial resolution of voice rendering, as compared to a perceptually realistic spatial rendering of voice. Thus, this model allows for predicting the quality of spatial rendering of voice depending on the computing capacity, i.e. the SH order intended to be achieved. This is particularly useful in the framework of virtual acoustics, for applications such as telepresence, visioconferencing, or video gaming, where the balance between perceptual accuracy and computing cost is always a hot topic.

4 Conclusion

In this pilot experiment, a perceptual threshold for the spatial resolution of the human voice directivity in different situations was determined. The threshold value averaged across all scenarios is SH order 8.4, which is slightly higher than the perceptual threshold found in previous research [9]. The variable parameters (voice type, head orientation, and room acoustics) investigated in this listening test showed specific trends which should be confirmed by further studies involving more participants and scenarios. The results underlined the perceptual relevance of high frequency in speech rendering in the context of sound spatialization. Further research include more detailed investigation of the different parameters that might influence spatial speech perception by means of advanced statistical methods such as the linear mixed-effects models. In addition, an accurate model of similarity was proposed. It showed the logarithmic nature of the relationship between SH order and ratings of similarity to very high spatial resolution. Numerous applications in virtual reality will benefit from these findings because the trade-off between computing cost and perceptual accuracy remains a central issue.

Acknowledgments

The authors would like to thank the participants of the listening test. This work was supported by a grant of the Alexander von Humboldt Foundation.

Conflict of interest

The authors declare they have no conflicts of interest in relation to this article.

References

1. F. Trendelenburg: Beitrag zur Frage der Stimmrichtwirkung [Contribution to the question of the directivity of voice]. Zeitschrift für technische Physik 1, 11 (1929) 558–563.
2. P. Kocon, B.B. Monson: Horizontal directivity patterns differ between vowels extracted from running speech. Journal of the Acoustical Society of America 144, 1 (2018) EL7–EL12.
3. W.T. Chu, A.C.C. Warnock: Detailed directivity of sound fields around human talkers. NRC-CNRC, NRC Publications Archive Archives des publications du CNRC, 2002, pp. 1–47.
4. B. Katz, C. d'Alessandro: Directivity measurements of the singing voice, in 19th International Congress on Acoustics, Madrid, Spain, 2–7 Sept, 2007.
5. J.L. Flanagan: Analog measurements of sound radiation from the mouth. Journal of the Acoustical Society of America 32 (1960) 1613–1620.
6. J. Struve: Directivity of real and artificial speakers for room acoustic measurements. Master's Thesis, Tech. Univ. Berlin, 2018.
7. B.B. Monson, E.J. Hunter, A.J. Lotto, B.H. Story: The perceptual significance of high-frequency energy in the human voice. Frontiers in Psychology 5 (2014) 587.
8. J. Wendt, B. Weyers, J. Stienen, A. Bönsch, M. Vorländer, T.W. Kuhlen: Influence of directivity on the perception of embodied conversational agents' speech, in Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, Paris, France on 2–5 July, 2019. ACM. 2019, pp. 130–132.
9. M. Frank, M. Brandner: Perceptual evaluation of spatial resolution in directivity patterns, in DAGA Conf., Rostock, Germany, 18–21 March 2019.
10. A. Lindau, S. Weinzierl: An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom, in VDT Int. Conv., Leipzig, Germany on 16–19 Nov, 2006.
11. ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems. International Telecommunication Union, Geneva, Switzerland, 2015.
12. S. Le Bagousse: Elaboration d'une méthode de test pour l'évaluation subjective de la qualité des sons spatialisés [Elaboration of a test method for the subjective evaluation of the quality of spatialized sounds]. Acoustique, Univ. Bretagne Occidentale, Brest, France, 2014.
13. C. Porschmann, J.M. Arend: A method for spatial upsampling of voice directivity by directional equalization. Journal of the Audio Engineering Society 68, 9 (2020) 649–663.
14. B. Rafaely: Fundamentals of spherical array processing, Vol. 8. Springer, 2015.
15. M. Schoeffler, F.-R. Stoter, B. Edler, J. Herre: Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA), in 1st Web Audio Conf., Paris, France on 26–28 Jan, 2015, pp. 1–6.
16. J.F. Culling, Q. Summerfield, D.H. Marshall: Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. Speech Communication 14, 1 (1994) 71–95.

Cite this article as: Quélennec A. & Luizard P. 2022. Pilot study on the influence of spatial resolution of human voice directivity on speech perception. Acta Acustica, **6**, 10.