


Binaural Recordings in Natural Acoustic Environments: Estimates of Speech-Likeness and Interaural Parameters

Trends in Hearing
Volume 24: 1–19
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2331216520972858
journals.sagepub.com/home/tia


S. Theo Goverts¹  and H. Steven Colburn²

Abstract

Binaural acoustic recordings were made in multiple natural environments, which were chosen to be similar to those reported to be difficult for listeners with impaired hearing. These environments include natural conversations that take place in the presence of other sound sources as found in restaurants, walking or biking in the city, and so on. Sounds from these environments were recorded binaurally with in-the-ear microphones and were analyzed with respect to speech-likeness measures and interaural difference measures. The speech-likeness measures were based on amplitude–modulation patterns within frequency bands and were estimated for 1-s time-slices. The interaural difference measures included interaural coherence, interaural time difference, and interaural level difference, which were estimated for time-slices of 20-ms duration. These binaural measures were documented for one-fourth-octave frequency bands centered at 500 Hz and for the envelopes of one-fourth-octave bands centered at 2000 Hz. For comparison purposes, the same speech-likeness and interaural difference measures were computed for a set of virtual recordings that mimic typical clinical test configurations. These virtual recordings were created by filtering anechoic waveforms with available head-related transfer functions and combining them to create multiple source combinations. Overall, the speech-likeness results show large variability within and between environments, and they demonstrate the importance of having information from both ears available. Furthermore, the interaural parameter results show that the natural recordings contain a relatively small proportion of time-slices with high coherence compared with the virtual recordings; however, when present, binaural cues might be used for selecting intervals with good speech intelligibility for individual sources.

Keywords

everyday-life recordings, binaural recordings, speech-likeness, interaural differences, natural environments

Received 25 November 2019; Revised 9 October 2020; accepted 15 October 2020

Participation in social interactions is an important priority for listeners with impaired hearing; hence, an important goal of clinical audiology is to optimize auditory function in multiple-source environments. Speech recognition in these environments is a major component of auditory function in the context of social interactions, and a lot of research has been done in this area including the role of binaural processing in environments with multiple sources and/or reverberation (e.g., Best et al., 2017; Beutelmann et al., 2010; Bronkhorst, 2000; Culling et al., 2004; Hawley et al., 2004; Kidd et al., 2019; Lavandier & Culling, 2010; Marrone et al., 2008a,

¹Otolaryngology-Head and Neck Surgery, Ear & Hearing, Amsterdam Public Health, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

²Biomedical Engineering Department, Boston University, Boston, Massachusetts, United States

Corresponding author:

S. Theo Goverts, Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology-Head and Neck Surgery, Ear & Hearing, Amsterdam Public Health, Amsterdam, Netherlands. PO Box 7057, Amsterdam, 1007 MB, The Netherlands.

Email: st.goverts@amsterdamumc.nl



2008b, 2008c). In most of these studies, the multiple-source environments were constructed from idealized sources, sometimes with reverberation added, but always appropriately controlled to allow explicit modeling and parameter isolation. These conditions allow application of theoretical models and they can illustrate and test models and hypotheses, but they are also simpler than most of the environments in which we normally operate in everyday life. This study recorded and analyzed waveforms from naturally occurring multi-source environments and these recordings are available for research use upon request.

Recently, researchers have started to investigate the acoustics of everyday environments more deeply. For example, Smeds et al. (2015) analyzed bilateral signals that had been recorded in various environments using microphones on a headband as described by Wagener et al. (2008). Smeds et al. analyzed these recordings with respect to noise levels and estimated signal-to-noise ratios (SNRs) with the use of a manual noise estimation procedure in cases when there were intervals containing speech in the various environments measured. Their analyses focused on estimating the SNRs in naturally encountered environments. Their results showed that there was a wide range of naturally occurring SNRs, estimated monaurally, and that these SNRs are usually above zero decibels. The A-weighted SNRs show a rather flat distribution varying between -4 and 34 dB and between -10 and 30 dB for best and worse ear, respectively. In another study, Wu et al. (2018) analyzed the speech level, noise level, and SNR of 894 listening situations that were recorded by 20 older adults who had mild-to-moderate hearing loss and who carried single-channel digital recorders. This was accompanied by *in situ* surveys on smartphones several times per day to report the characteristics of their current environments. When speech listening was indicated (judged subjectively) by the participants, SNR was estimated manually. The results showed a more peaked distribution with SNRs ranging from -10 to 30 dB in these listening situations. The majority of SNRs were between 2 and 14 dB; only 7.5% of the situations were very noisy (SNR <0 dB). Weisser & Buchholz (2019) studied conversational SNRs in realistic conditions. They elicited conversations between pairs of subjects at fixed positions who listened to binaural realistic recordings during the conversations. By recording voices and analyzing speech levels as a function of the level of the background recordings, they found that realistic SNRs vary between -10 and $+15$ dB. Negative SNRs occurred only for background levels above 69 dB SPL at a fixed-position distance of 1.0 m or above 75 dB SPL at a fixed position distance of 0.5 m.

Interaural difference parameters in realistic scenarios were studied by Mlynarski and Jost (2015). Their study

focused on binaural recordings made in three natural environments, and they analyzed the interaural time differences and interaural intensity differences in narrowly tuned (i.e., one-third octave) frequency channels. The acoustic environments included recordings that are descriptively called “nocturnal nature,” “forest walk,” and “city center.” They found that, in situations with multiple sources, the overall distributions of interaural differences are not easily separated into distinct sources since the combinations, echoes, and reverberations lead to interaural differences that do not match pure single sources.

In the research reported here, bilateral acoustic signals in a number of realistic environments were recorded and analyzed. This allows a description of the nature of the stimuli that reach the two ears in everyday situations, sometimes referred to as the “bilateral vibration pattern” (e.g., Bregman, 1990). These stimuli are also, of course, the inputs to hearing-assist devices (hearing aids, cochlear implants, etc.). The analysis of these bilateral stimuli allows us to investigate the nature of the information that is available to the binaural hearing system in realistic scenarios. Furthermore, the recordings provide data about the so-called better-ear effect by comparing the speech-likeness (SL) at the two ears (cf. Cosentino et al., 2014). In clinical audiology, this better-ear effect has served as an argument for bilateral hearing rehabilitation with hearing aids or cochlear implants (e.g., Culling et al., 2012). One important question is whether this better-ear effect is functionally important: In how many instances in real life is this advantage actually there and over what time intervals. In addition, the availability of binaural cues that facilitate true spatial listening (e.g., localization, binaural unmasking, and location-based segregation) can also be measured. Finally, the results can be compared with the characteristics of test configurations that are used in the clinic and in research.

The specific goals of this study were as follows:

- To make bilateral recordings in realistic listening scenarios that are relevant to daily life;
- To characterize these recordings: in particular, (a) how “speech-like” are the recordings and (b) to what extent do the recordings contain interaurally coherent information that may allow for the binaural system to enhance speech recognition;
- To investigate the functional importance of the better-ear effect, by comparing SL between the ears as an illustration of the type of clinically relevant research questions that can be addressed using such recordings; and
- To compare parametric properties of natural recordings to virtual recordings that mimic conditions that are used in regular clinical testing of speech recognition.

Natural recordings were made using two commercially available microphones, one in each concha, that were connected to a portable data-recorder. Recordings were made for two types of environments, natural environments and virtual environments. The natural-environment recordings were made in a variety of locations that included Inside (e.g., in a home or restaurant), Outside (e.g., a walk in the city), and Public Transport (e.g., in a bus) environments. The virtual-environment recordings were constructed computationally from speech targets combined with maskers with varying degrees of similarity to speech. The virtual locations of speech and masker waveforms were also varied to create recordings with different spatial characteristics, and reverberation was included. All of the recordings, both natural and virtual, were processed to characterize their SL and their binaural properties. The binaural characteristics were estimated for short time-slices in selected frequency bands. It should be noted that speech-likeness is used as a surrogate for speech intelligibility because the latter is difficult to estimate automatically with current methods.

The processing used to quantify the SL of the recorded signals is nonintrusive; it does not require knowledge of the speech content (and thereby avoids privacy issues) and does not use a priori knowledge about target speech and masker. The SL approach used an extended version of the Speech Intelligibility Index (SII) and the Speech Transmission Index. The approach used here is based on the distribution of modulation strength in several frequency bands compared with typical patterns obtained for natural speech in a quiet environment (e.g., Dubbelboer & Houtgast, 2008; Houtgast & Steeneken, 1972; Jørgensen & Dau, 2011; Jørgensen et al., 2013). The most general approaches of this type are based on a speech-relevant modulation pattern. Our specific approach uses a comparison of low-frequency modulation strength to higher frequency modulation strength, leading to a measure of SL. This measure, defined in detail later, was used by Falk et al. (2010) and Cosentino et al. (2014).

To investigate the availability of useful binaural information, three binaural parameters were measured for each 20-ms time-slice within each of two frequency bands. Specifically, three quantities were computed: interaural cross-coherence (ICC), defined as the normalized cross-correlation at the interaural time delay (ITD) with maximum correlation, the ITD giving this maximum, and the interaural level difference (ILD). These quantities were computed for several time-slice durations for low (500 Hz) and high (2000 Hz) frequency bands (one-quarter octave). In the analysis of the high-frequency band at 2 kHz, all interaural measures (ICC, ITD, and ILD) were computed using the envelopes of the bandpassed signals.

All measures of SL and interaural differences were computed for both the natural and the virtual recordings so that measures in natural environments could be interpreted with respect to values estimated from the virtual recordings. All of these results are discussed with respect to the evaluation of hearing abilities.

Methods

Natural Recordings

As noted earlier, the goal of recording stimuli that naturally occur in complex sound environments was to provide a library of binaural stimuli that would (a) be available for experiments and analysis and (b) provide a resource for understanding the difficulties experienced by hearing-impaired listeners, even when they are provided with high-quality hearing devices. To determine what sound environments and stimuli would be appropriate for this resource, we interviewed hearing-impaired patients and consulted with experts in audiology and otolaryngology, both in the United States and in The Netherlands. On the basis of these interviews, we selected a variety of environments and recorded pressures in the ear canals with binaural microphones. The selected environments included conversations walking or biking through town, conversations at a table in a restaurant, conversations in a typical home environment, sounds riding on a bus, and sounds in a large railroad station with announcements and other typical sounds. These two-channel recordings allow listening and subjective evaluations as well as analysis of the characteristics of the recorded stimuli. The recordings sound realistic when listened to with appropriate binaural headphones. In Table 1, the characteristics of eight distinctive environments in which natural recordings were made are summarized, descriptively referred to as Home, Restaurant, City Walk, City Talk, City Bike, Station Hall, In a Train, and In a Bus. The digital recorder used was an Olympus linear PCM recorder (Model LS-11), allowing recordings of significant length (potentially multiple hours) and of high quality (viz., a sample rate of 44,100 samples per second and a level resolution of 16 bits per sample). Specifically, measurements used commercially available microphones (Sound Professionals MS-TFB-2) with recording durations varying from tens to hundreds of seconds. These recordings are available upon request. Although other recordings were made, only these tabulated recordings are available upon request. Note that recordings were made in environments in which the human-subject participants normally interacted and that the subjects were fully informed about the use of the recordings. The study was approved by the Medical Ethics Committee of VU University Medical Center (Amsterdam).

Table 1. Summary of Natural Environments Measured in This Study.

Environment	Situation	Other sound sources
HOME	Two people conversing at home in living room; male wearing mics; female partner.	Sounds of having dinner; walking to kitchen, occasional kitchen noise; radio music switched on.
RESTAURANT	Two people having brunch in small restaurant; female wearing mics; male dinner partner.	Table with two females talking to the right (approx. 1.5 m); kitchen to the left/back (approx. 3 m).
CITY WALK	Two people walking through city streets; female wearing mics; male partner.	People passing, cars passing, wind, changing direction, more people passing, quiet area.
CITY TALK	Two people sitting at a fixed location in the inner city; female wearing mics; male partner.	Voices, street music, cars, motor scooters passing, children yelling.
CITY BIKE	Two people making a bike ride through city; male wearing mics; female on a different bike.	Sounds include getting bikes unlocked; traffic noise; cars, motors, scooters; wind noise; traffic-light sounds.
STATION HALL	Listener with mics sitting in a Train station.	Noises of trains arriving and leaving, speech from nearby persons and from occasional PA announcements.
IN A TRAIN	Listener with mics sitting in train, window on left side.	Sounds from multiple travelers who are talking and making phone calls, broadcast announcements.
IN A BUS	Listener with mics sitting in a bus.	Multiple stops; with short public address (PA) announcements, hardly intelligible.

Virtual Recordings

In addition to the natural recordings described in Table 1, virtual recordings of simple environments were created and used for reference and comparisons. The virtual recordings were created by starting with single-channel anechoic waveforms (speech or noise) and then processing these waveforms using head-related transfer functions to create binaural signals as would be generated from various locations in the horizontal plane in well-defined spaces. Spatial conditions (and associated abbreviations) included target and masker colocated straight ahead at 0° (T0M0); target at 0°, and masker at +45° to the right, respectively (T0M+45); target left at -45° and masker at +45° (T-45M+45); and target at 0° with two maskers, one at +45° and one at -45° (T0M+/-45). The target stimulus was always speech, a grammatically correct sentence consisting of about eight or nine syllables, from a single female talker (Versfeld et al., 2000) and the masker was selected for each condition from three options: (a) a Gaussian noise with long-term-average-speech-shaped spectrum (LTASS noise); (b) a fluctuating (speech-envelope-modulated) noise (FLUC), created by modulating the envelope of the LTASS noise with an envelope from the female speech as described by Festen and Plomp (1990; FLUC noise); or (c) a male talker (MALE) (grammatically correct sentences consisting of about eight to nine syllables; MALE noise). Virtual stimuli were computed for both anechoic conditions (no reflections) and reverberant conditions via a (rectangular) room model developed by Joseph Desloge (e.g., Shinn-Cunningham et al., 2001). In our computations, the model simulates a room with dimensions 10 × 10 × 4 m and a source-to-

head separation of 1.41 m. In the reverberant simulations, reflections were added with an absorption coefficient of 0.14 for all six surfaces of the virtual room, resulting in a T60 reverberation time of about 1.3 s (calculated from the Sabine formula, e.g., Sabine, 1900/1915).

Measure for Characterizing SL of the Recordings

A measure of SL was calculated for each of six octave bands and then combined over bands with a SII weighting for each band in the combination. The SL measure for each band was based on its modulation strength versus modulation frequency, specifically, on the strengths of low modulation frequencies relative to the strengths at high modulation frequencies. This modulation-strength measure was developed and used by Falk et al. (e.g., Cosentino et al., 2014; Falk et al. 2010) in their studies of speech quality and intelligibility. Specifically, low modulation frequencies were considered to be related to speech and high modulation frequencies were considered to be related to noise or reverberation. Therefore, the SL in each frequency band was defined in terms of the logarithm of the ratio between the mean modulation energy at low modulation frequencies (4, 8, and 16 Hz) and the mean modulation energy at high modulation frequencies (32, 64, and 128 Hz).

The details of the method for estimating the SL parameter, which is abbreviated here as SL, are as follows:

1. Divide the ear signal being processed into 1-s time-slices with 0.5-s cosine-shaped rise and fall times with no overlap of windows.

2. Apply bandpass filters to the waveform, resulting in six, one-octave frequency bands f_i where f_i values are centered at 250, 500, 1000, 2000, 4000, and 8000 Hz.
3. Extract the envelope of each frequency band using Hilbert transforms.
4. Calculate the Discrete Fourier Transform of the envelope for each frequency band.
5. Determine, for each frequency band, the modulation energy in each of six one-octave modulation bands centered at 4, 8, 16, 32, 64, and 128 Hz by taking the sum of the squared Discrete Fourier Transforms in each band.
6. Calculate, for each f_i , the ratio between (a) the average of the modulation energy in the modulation frequency bands centered at 4, 8, and 16 Hz and (b) the average of the modulation energy in modulation frequency bands centered at 32, 64, and 128 Hz.
7. Take 10 times the log (base 10) of this ratio to get the speech-likeness $SL(f_i)$ for each f_i .
8. Using the SII frequency weighting (indicated as SII(f_i)) as weights for the SL values, take the SII-weighted sum of the SL values across f_i .

Representing these steps with equations, letting $SL(f_i)$ represent the SL for each f_i and SL to represent the weighted average across FBs, we have

$SL(f_i) = 10 \log_{10} \{ \text{mean}(\text{mod strength}[4 \text{ to } 16 \text{ Hz}]) / \text{mean}(\text{mod strength}[32 \text{ to } 128 \text{ Hz}]) \}$ and then finally

$$SL = \sum_{i=1}^6 SII(f_i) * SL(f_i).$$

We checked the usefulness of this SL metric by applying it to some nonspeech sounds from the ICRA Natural Sound Library (<https://icra-audiology.org/Repository/icra-noise>) and to samples of natural speech from a male and a female talker recorded with our binaural microphones (described earlier). The data are presented in Figure 1 with SL values for each 10-s time interval. Results show a nonoverlap between values of SL for speech sounds and for nonspeech sounds and further indicate no gender differences for speech signals.

Measures to Investigate the Interaural Difference Information Available in the Recordings

For the interaural parameter estimates, the left and right signals were filtered through a bank of quarter-octave-bandwidth filters, and selected frequencies were analyzed with respect to their interaural differences. In the results reported here, one low-frequency band (centered at 500 Hz) and one higher frequency band (centered at 2000 Hz) were chosen as representative frequencies for binaurally distinctive subdivisions. Also, since it is generally believed that only envelope timing information is

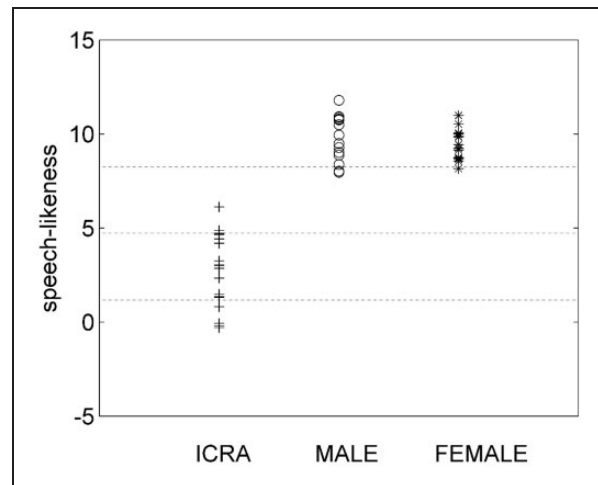


Figure 1. Values of speech-likeness for (A) a set of ICRA non-speech recordings downloaded from <https://icra-audiology.org/Repository/icra-noise>; (B) recordings using the binaural microphones of clean natural speech by 15 male talkers; and (C) recordings using the binaural microphones of clean natural speech by 15 female talkers. Dotted lines indicate boundary values corresponding to 0%, 50%, and 100% correct speech recognition (explained later in the Discussion section).

available at high frequencies, for the band at 2000 Hz, only the interaural information contained within the envelopes of the 2000-Hz waveforms was examined. In all cases, the bilateral signals were analyzed in 20-ms time-slices, with 5-ms, cosine-shaped rise and fall times, making a total (nonzero) window duration of 30 ms. Successive time-slices were shifted by 20-ms, so that there was overlap in the ramps, but the basic width and spacing were 20 ms. The 20-ms-based time-slice duration was chosen based on other studies using this duration (Best et al., 2017; Brungart et al., 2006).

Our binaural processing starts by computing the cross-correlation function, a classic component of most binaural models, for each time-limited and frequency-limited slice (called a time-frequency or TF slice). For each TF slice, the left and right waveforms are used to compute the cross-correlation function, which is normalized by the product of the root-mean-square (rms) amplitudes, giving a maximum value of unity for waveforms that differ only in a pure delay and/or a fixed scale factor. The time shift for which the cross-correlation function reaches its maximum value is identified as the ITD for this TF slice, and the maximum normalized correlation is defined as the ICC. If the ICC of a certain TF slice was less than 0.5, then that slice was regarded as unreliable and hence eliminated from the statistics for ITDs. Also, ITD values were constrained to fall in the “natural range” within 1 ms of zero delay; that is, slices with maximum ITD magnitudes that were larger than 1 ms were eliminated from the ITD distributions. The ILD

of the slice is defined as the decibel difference of the rms amplitudes (i.e., 20 times the base-ten logarithm of the rms ratio). The ILD was estimated for every time-slice, and all values of ILD were included. The fraction of rejected ITDs depended on the conditions of the recording and varied between 23% (for the HOME recordings) and 37% (for the IN A BUS recordings). In some cases, other restrictions were considered and applied such as ICC values that were considered highly coherent with cutoff of 0.95 instead of 0.5; these cases are noted and discussed further later. Note also that the envelope waveforms, being nonnegative, always have nonnegative correlation values (for every ITD), and so the ICCs for 500 Hz and 2000 Hz cases are not directly comparable (cf. Bernstein & Trahiotis, 1996a, 1996b).

Results

SL results are presented and discussed before the interaural parameter results. Within each type of results, parameter values calculated from the analyses of the virtual stimuli are presented first. The distributions of these parameters are then used to interpret the corresponding values computed from the natural recordings.

Speech-Likeness

SL Measures in Virtual Recordings. SL values for two target and masker placements (T0M0 and T-45M+45) out of the four described earlier (T0M0, T0M+45, T-45M+45, and T0M+/-45) are shown in Figure 2 as a function of the target-to-masker ratio (TMR), defined by the ratio of the target and masker levels at the speaker locations

(in decibels). The T0M0 case (both sources straight ahead) is shown in the first column and T-45M+45 (target to the left and masker to the right) is shown in the second column. The upper row gives the data for the anechoic condition; the lower row is for the reverberant condition. In each panel, the dependence of SL on TMR is shown separately for each ear for each of the three different masker types (LTASS, FLUC, and MALE), with the six cases coded by symbols (as given in each panel). For the anechoic condition, when the speech dominates (e.g., when the TMR approaches or exceeds about 25 dB), all SL values are similar ($SL \approx 12.5$). Note that, as expected for these well-pronounced speech materials, the SL values are in the upper range of values obtained for the clean natural speech (see Figure 1). As TMR decreases, the different maskers show different SL values.

The pattern of results for the colocated case (T0M0, first panel) shows, as expected, no visible differences between left and right sides. Of course, for an actual human being, as opposed to our simulations with symmetrical head-related transfer functions, the ears are not perfectly identical, so some left-right asymmetries would be expected in real life. Considering the MALE masker first (diamonds), we see that, for large TMR values when the FEMALE target dominates, the SL is about 12.5 and for small TMR values (say -30 dB) when the MALE masker dominates, the SL is about 11, showing that the SL varies between voices. From Figure 1, we see that the male and female distributions are similar relative to the interindividual listener variability. When the LTASS masker is used, the low-TMR values of SL are

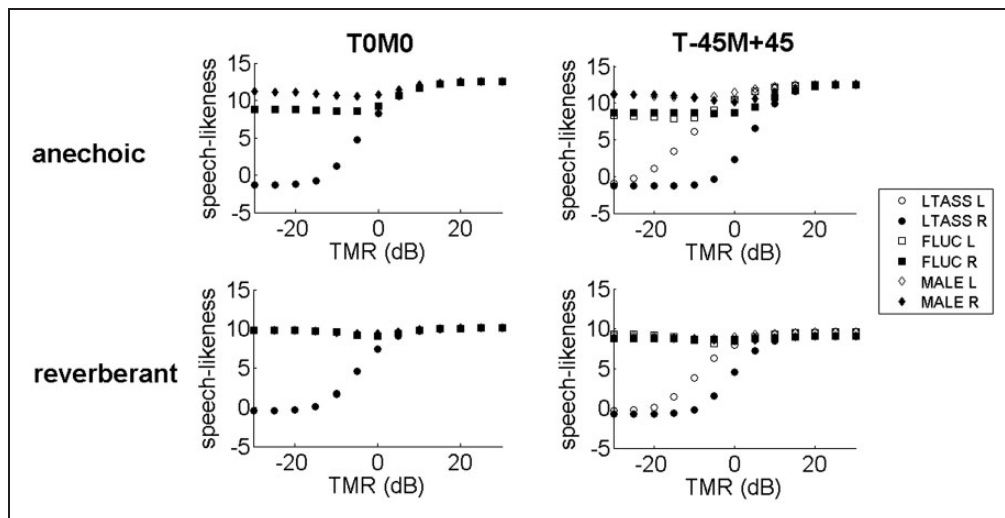


Figure 2. Measures of Speech-Likeness in the Virtual Recordings as a function of the Target-to-Masker Ratio (TMR), for three Masker Types (LTASS, FLUC, and MALE) and For each ear (L,R) as Noted by Different Symbols. The top row is for anechoic conditions: target and masker colocated at 0 (left column) and target at -45 and masker at +45 (right column). The second row presents the same configurations for a highly reverberant condition ($T60 > 1.2$ s). LTASS = long-term-average-speech-shaped spectrum. FLUC = fluctuating (speech-envelope-modulated) noise, MALE = a male talker; see text in methods section.

only about -1.3 , consistent with the lack of SL in the modulations of the LTASS noise. Finally, as expected, the SL measure for the FLUC masker is close to the MALE case, with masker-dominated cases reduced slightly, from 11 to 9. Note that the SL measure strongly discriminates target speech from the LTASS noise masker but discriminates less between target speech and the FLUC and MALE maskers. Apparently, the FLUC noise contains roughly the same amount of speech-relevant modulations as speech, as was intended in its design (cf. Festen & Plomp, 1990), although the original noise envelope results in a final FLUC envelope that is more irregular than the envelope of pure speech.

Considering now the nonsymmetric case ($T-45M+45$) in the right panel of the upper row, clear left-right differences are observed with the LTASS masker. The left-right differences in SL are almost 10 SL units when the TMR is near -5 dB. Left-right differences in the FLUC and MALE maskers are much smaller, as expected since these maskers are also very speech-like. Finally, we note that the $T0M+45$ condition (not shown) has similar patterns in responses, as one would expect with the similar but smaller asymmetries. For the other condition not shown, the three-source condition ($T0M+/-45$), one observes a similar overall dependence on SNR with no left-right differences, as expected from the statistical symmetry in this case.

The reverberated conditions in the lower row of Figure 2 show, again as expected, reduced SL values at high TMRs (reduced from about 12.5 to about 10.0). This reduction also holds at low TMR values for the MALE masker. In fact, the difference between MALE and FLUC masker disappears in the reverberated conditions. The interaural differences in SL are also reduced with reverberation in the $T0M+45$ and $T-45M45$ conditions, reflecting the reduction of the better-ear effect in reverberant environments.

Because we are specifically interested in the better-ear effect, which is related to interaural differences in SL, a reference panel for comparing data between the ears is given in Figure 3 for some of the virtual stimuli. Here, SL values of the right ear are plotted versus the left-ear values as the TMR varies for various spatial configurations for the LTASS masker, since these cases show the largest interaural differences (cf. Figure 2). SL values for anechoic cases are plotted as solid black curves and SL values for reverberant cases are shown as gray curves. Specifically, the values plotted come from computed values of SL, some of which are plotted in Figure 2: The configurations are $T0M0$ (along the diagonal, by symmetry), $T-45M+45$, and $T45M-45$ (labeled curves, using data as plotted in the upper [anechoic] and lower [reverberant] right panels of Figure 2, together with the assumed left-right symmetry). Finally, the aforementioned SL values span a broad range of values as

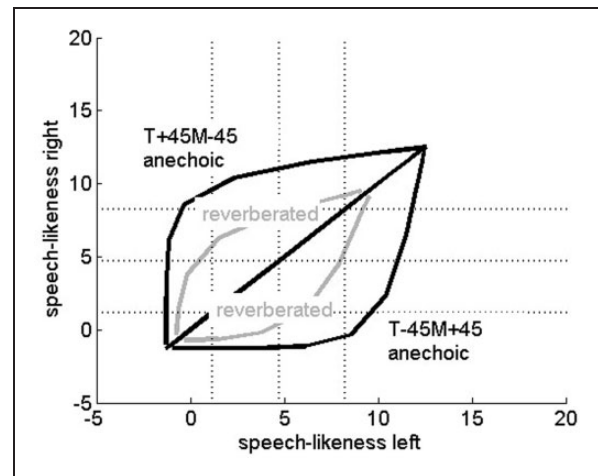


Figure 3. Reference Frame for Interpreting Speech-Likeness (SL) in the Natural Recordings. Calculations for the values plotted use virtual stimuli and are explained in the text. For interpreting interaural differences in the natural recordings, SL data for the right ear are plotted versus SL data for the left ear for the $T-45M+45$ condition in LTASS noise (and for the mirrored condition $T+45M-45$). Data for the anechoic conditions are shown with dark lines and data for the reverberated cases are given with gray lines. Boundary values for speech-likeness values roughly corresponding to 0%, 50%, and 100% correct speech recognition are indicated by dashed horizontal and vertical lines, as described in the Discussion section.

represented in Figure 1. From an overall perspective, the data in Figure 3 show that large interaural differences in SL values can occur. For example, in the $T-45M45$ or $T45M-45$ configuration, SL can be less than 1.2 (corresponding to typical values of SL for nonspeech ICRA recordings as seen in Figure 1) for one ear and be greater than 8.2 (corresponding to pure speech for speakers in Figure 1) for the other ear. Differences are smaller for reverberated cases.

SL Measures in Natural Recordings: Exploration of the Data. Measures for SL in natural recordings are given in Figure 4A to H. For each of the eight environments, the data are presented in the two panels of the indicated subfigure, with Figure 4A corresponding to the HOME environment, Figure 4B corresponding to the RESTAURANT environment, and so on, as ordered in Table 1. In each subfigure, the left panel shows each of the two SL measures plotted as a function of time for a 250-s portion of each recording, and the right panel shows the SL for the right ear plotted versus SL for the left ear (as shown in Figure 3 for the virtual recordings). The overall picture that emerges in the SL data is a large variance within and between environments, both in the interaural differences and in their temporal dynamics.

The SL *distributions* can be extracted visually from left-panel plots, and the distributions within each of the three

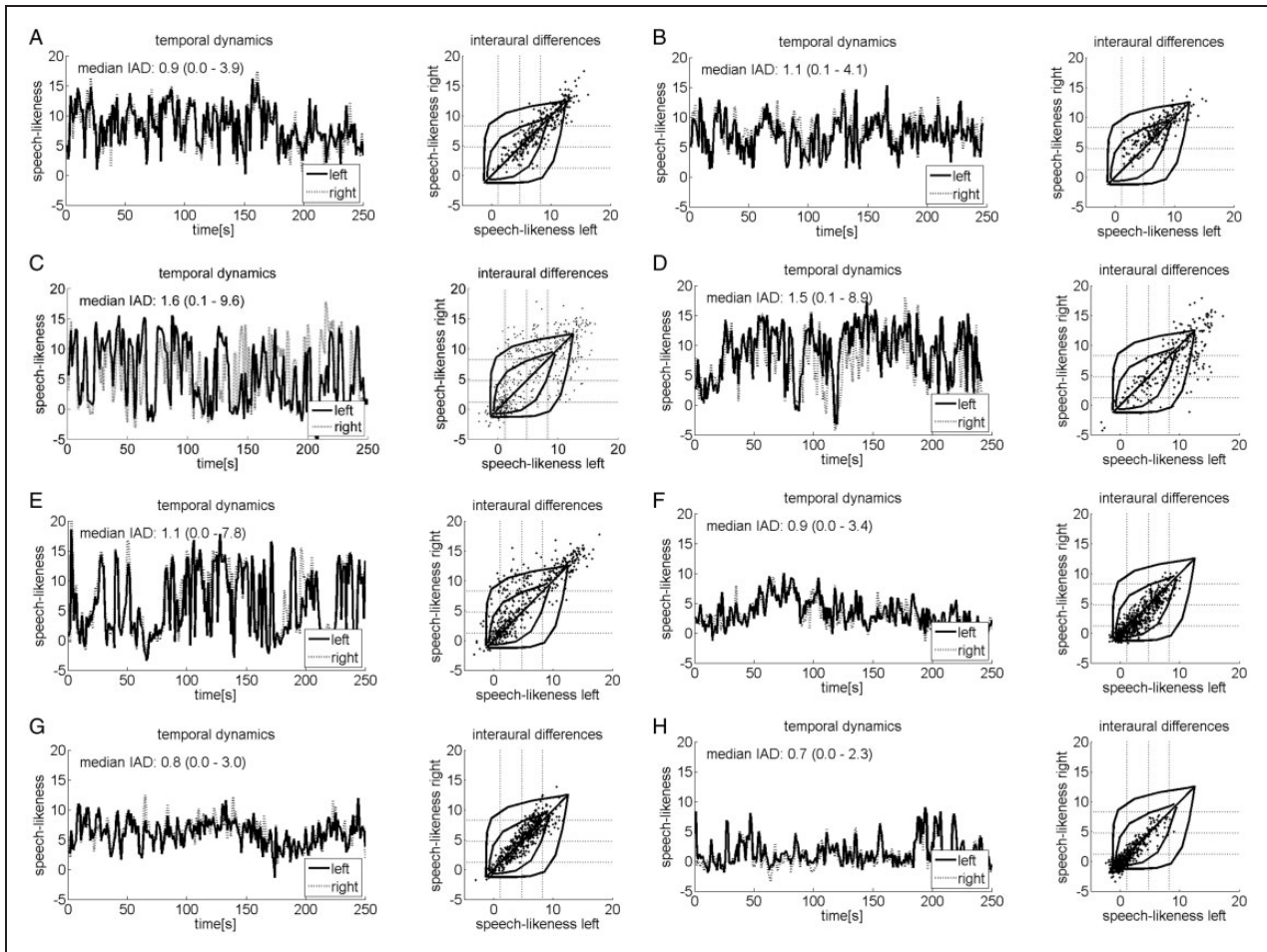


Figure 4. Speech-Likeness (SL) in Natural Recordings for Two Inside Environments [Home (a) and Restaurant (b)], for Three Outside Environments [City Walk (c), City Talk (d), and City Bike (e)], and for Three Public Transport Environments [Station Hall (f), in a Train (g), and in a Bus (h)]. For each environment, the left panel illustrates the temporal dynamics of SL in natural settings by plotting speech-likeness values for both ears as a function of time for a 250-s portion of each recording; median values of the interaural differences (IADs) are given in each plot as well as the boundaries of the 95% intervals. The right panel gives interaural differences in speech-likeness by plotting values for the right ear versus those for the left ear in the reference frame provided by Figure 3.

categories of environments show similarities. For the Inside environments, both Home and Restaurant (Figure 4A and B) have median values around 7 or 8 at both ears (corresponding to SNRs around 0 dB in speech spectrum noise [LTASS]; see Figure 2 earlier) and the 95% interval ranges from 2 to 14 (roughly). For the Outside environments in Figure 4C to E (City Walk, City Talk, and City Bike), one sees broader distributions at both ears that are sometimes bimodal (City Walk and City Bike) with the 95% interval ranging from -1 to 15. For the Public Transport environments in Figure 4F to H (Station Hall, In a Train, and In a Bus), one sees the lowest values, most values in the range 0 to 5 (corresponding to SNRs in LTASS around -10 to -5 dB in Figure 2) with the 95% interval ranging from -1 to 8.

The highest *interaural differences* in SL (as seen in the deviations from the diagonal in the right panel) are

found for the Outside environments (Figure 4C–E), most prominently for the City Walk, where interaural differences exceed the T -45 M $+45$ reference values, for example, where the signal at one ear is a nonspeech-like signal and the other is very speech-like. Also, the inside environments (Figure 4A and B) have slices with smaller but substantial interaural differences. Interaural differences are smallest for the samples of Public Transport recordings that are included in Figure 4F to H. The recordings in Restaurant, City Bike, and City Talk show a clear overall dominance of SL for one ear (right, right, and left, respectively) in accordance with the configuration of sounds in those settings and with perceptions listening to these recordings.

Considering the temporal fluctuations seen in the left panels of Figure 4, the degree of fluctuation in the SL over different environments roughly concur with the

pattern of interaural differences. The largest fluctuations are found for the Outside environments and lowest for the Public Transport environments, consistent with the movements of the dominant sources as perceived in listening to the recordings.

In Figure 5, the SL values for the right ear versus the left ear are plotted, combined over all natural environments, again using the reference frame presented in Figure 3. This gives a rough indication of the variety of interaural differences in SL a person could encounter in everyday life environments. The combined values show significant left–right differences, as already noted in Figure 4. The curves showing the virtual interaural results are again included, as in Figures 3 and 4. In addition, horizontal and vertical lines of various types are included for comparison to expectations from measurements of speech intelligibility for LTASS maskers. These lines and comments about comparisons of the current data to these lines are included below in the Discussion section.

Interaural Difference Information

In the following subsections, distributions of interaural parameters obtained from TF slices of virtual recordings are presented and discussed. Then, these distributions are compared with distributions for the natural stimulus

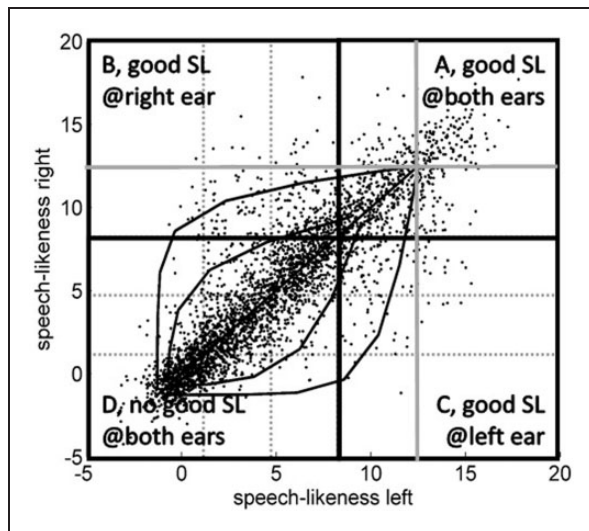


Figure 5. Speech-Likeness Values for Right Ear Versus Left Ear for all eight environments plotted together. As described in the Discussion section, four subdivisions are indicated for listeners with normal hearing, using the speech-likeness boundary of 8.2 corresponding to the SNR for 100% intelligibility in this population. These SL boundaries are indicated with thick black lines. In addition, boundaries for listeners with hearing impairment are indicated with gray lines. With impairment, the SNR for 50% intelligibility is shifted and the psychometric curve is shallower; hence, the SNR for 100% intelligibility will shift and so will the SL boundaries (toward higher speech-likeness values of about 11).

recordings. The virtual conditions presented here include a female target and a spatially separated male masker in anechoic space. The effects of adding reverberation are also illustrated. Multiple spatial conditions were processed, but they show patterns that are more or less as expected after considering the T–45M+45 cases, and so discussion here is focused on these T–45M+45 cases.

Interaural Parameter Distributions in the Virtual Recordings.

The interaural parameter distributions for the T–45M+45 cases for two masker–environment combinations, ANECHOIC and REVERBERANT, are shown in Figure 6A and 6B, respectively. In each subfigure, the leftmost column shows the distributions of ICC values for both frequency bands, the second column shows ITD versus ILD distributions that include all TF slices (with $ICC > 0.5$), and the third column shows these distributions including only the high-coherence TF slices (i.e., with $ICC > 0.95$). Note that, within each subfigure, the upper panels show the 500-Hz band values and the lower panels show the 2000-Hz band values. It was also observed that the ICC values for these two frequency bands appear as roughly independent in plots of their joint distributions (plots are not shown here). Data from other maskers and other spatial conditions are available and were processed, but they show patterns that are more or less as expected after considering the T–45M+45 cases shown. Also, note that all cases presented here use a TMR of 0 dB so that target and masker are equally strong in rms level at the source.

For the ANECHOIC masker (with no reverberation), shown in Figure 6A, the ICC distributions show mostly high values, many near unity; about 44% of the TF slices at 500 Hz have an ICC higher than 0.95; for the 2000 Hz envelope this fraction is similar at 43%. The ITD-vs-ILD plots show two broadly distributed clusters around the values we would expect for sources at $\pm 45^\circ$: Clusters near positive values roughly at 0.6 ms and 4 dB for 500 Hz and roughly 0.5 ms and 4 dB for 2000 Hz and with clusters near corresponding negative values, as expected because of the symmetric placements. Recall that only ITD data are shown from time-slices with ICC values above 0.50, so the ITD-vs-ILD plots are similarly restricted in the second column. As seen in the third column, the two distributions are more compact when the distributions are restricted to intervals with ICC values above the 0.95 criterion.

Turning to Figure 6B, for the REVERBERANT masker, we find a much broader distribution of ICC values at 500 Hz, with only 2% of TF slices with ICC higher than 0.95. The interaural difference distributions (ITD and ILD) are broader and nearly symmetric around zero ILD. This is not surprising since the direct

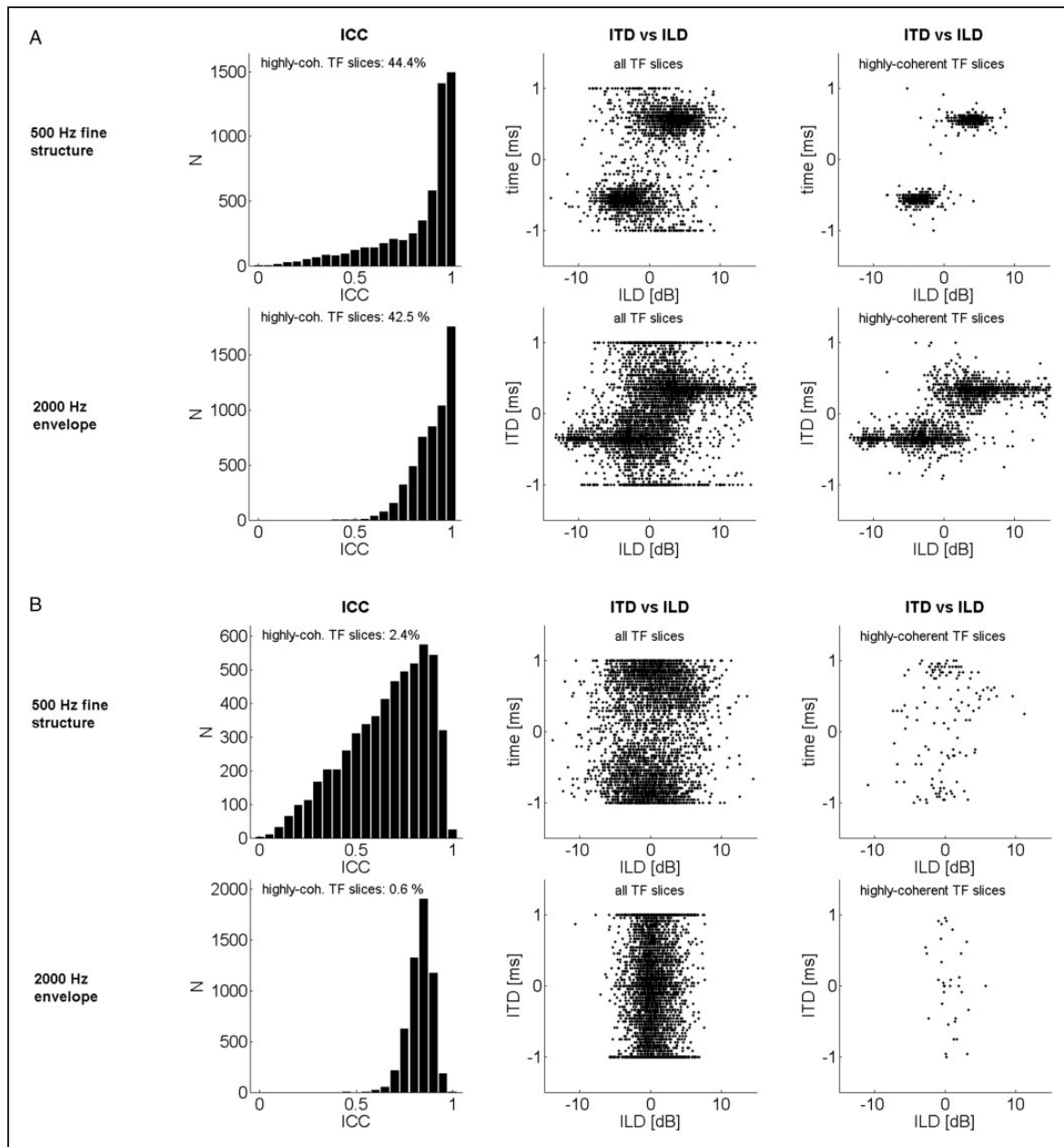


Figure 6. Interaural parameters analyzed in 20-ms time-slices for the virtual recordings for (A) ANECHOIC for spatial condition T-45M+45 and TMR = 0, and (B) REVERBERANT, also for spatial condition T-45M+45 and TMR = 0. Each subfigure presents in the top row (for bands centered at 500 Hz) from left to right: the distribution of ICC values; the joint distribution of ITD and ILD values using all ITD values recorded ($ICC > 0.5$), and corresponding ILDs; the joint distribution of ITD and ILD using only data with highly coherent ICCs ($ICC > 0.95$). In the second row, the same data are given for the envelopes of the 2000-Hz band. ICC = interaural cross-coherence; ITD = interaural time delay; ILD = interaural level difference.

sounds and the reflected sounds are combining to eliminate the dominance of the target and masker locations. Finally, note that the joint distributions of ITD and ILD change as expected when reverberation is added. The clear foci in the plots of Figure 6A, with one focus for each source, are replaced by a relatively diffuse pattern of points with minimal focus in Figure 6B, with only a relatively small amount of highly coherent TF

slices consistent with the interaural decorrelation of the waveforms in reverberation.

Although not shown here, other virtual environments with various combinations of masker types (MALE, LTASS, and FLUC) and spatial configurations (T0M0, T0M+45, T-45M+45, T0M+/-45) were evaluated and show expected patterns of parameter distributions, based on the cases discussed here.

Interaural Parameter Distributions in the Natural Recordings.

Recordings from the eight natural environments were processed to estimate the interaural parameters as described earlier for the same frequency bands and time-slice durations. Selected examples are presented in Figure 7A to C for Restaurant, City-Walk, and In-a-Train environments, providing one example each of Inside, Outside, and Public Transport environments. Specifically, each subfigure presents two rows of panels showing the interaural parameter samples, with 500 Hz in the upper row and 2000 Hz in the lower row. The panels in each row show the distribution of the ICC values for all time samples for each frequency, the temporal sequence of ILD values with ICC above 0.95, and the temporal sequence of ITD values for ICC above 0.95.

As an illustration of the interaural parameter distributions and their temporal variations, we consider the City Walk environment first, as displayed in the two rows of Figure 7B. The City Walk comprises recordings from the conchas of a female taking a walk outside in conversation with a male who accompanies her. As she walks, the environment includes sound sources from various directions in the form of voices of other walkers, children on bicycles, and buses driving by, as well as footsteps and other environmental sounds. It is clear in the leftmost panels that many time-slices have low ICC. In general, the ICC values are distributed broadly (even broader than the reverberant MALE masker condition in Figure 6B), consistent with the multiple sources and reflections that are present. For the interaural parameter versus time plots, values are estimated and presented, again only for slices with ICC values greater than 0.95. The ITD and ILD values are shown as time sequences of values to show temporal dynamics. It can be seen that the ILD and ITD values for 500 Hz show three distinct peaks: one near midline (both ITD and ILD near zero), and one on each side (broad ILD values near +5 dB [left side] and -5dB [right side], and ITD clusters near 0.6 ms [left side] and -0.6 ms [right side]). As seen in the temporal sequences of values, these clusters can be related to the movement of dominant sources. By listening to the recording it appears that the cluster of interaural differences near zero is generated by the voice of the female wearing the microphones and the side peaks are most often the male companion who is primarily on the left in the early part of the recording and more to the right in later parts of the recording.

Similar plots for the other natural environments, Restaurant and In a Train, are shown in Figure 7A and C, respectively. The distributions are generally consistent with the nature of the environments that one hears in listening to the recordings, with sources perceived in different directions and moving relative to the listener wearing the microphones. In the Restaurant

environment (Figure 7A), directions correspond to the dining partners, both near midline, and to additional sources off midline. These off-midline sources include people speaking at another table (to the right) and sounds from the kitchen direction (to the left). Overall, this Restaurant environment is dominated by the voices of the microphone wearer and her dining partner directly across the table, both of which provide minimal interaural differences. The negative ITD values in the time sequence are apparently due to people at a table to the right of the person wearing the microphones, corresponding to how the recording sounds when listening with headphones. In the temporal sequences in Figure 7C, there is a broad spatial distribution corresponding to sources to the right with high reverberation, consistent with riding in a Train and sitting on the left side of the car.

General Remarks on the Interaural Parameter Distributions in the Natural Recordings. A primary implication of the analysis of the virtual and natural recordings is that the perceptual system must be able to focus on high-coherence intervals and to extract information from relatively short time-slices in order to process information from multiple sound sources, although the specific limitations of time-slice sensitivity are not clear, as discussed later. This implication is supported by the subjective experience of listening to the recordings; though without formal measurements, these impressions have been clear to all who have listened to the recordings. This observation is consistent with the observations of Mlynarski and Jost (2015) based on their recordings from natural environments; when there are multiple sources, the overall distributions of interaural differences are not easily separated into distinct sources since the combinations, echoes, and reverberations lead to interaural differences that do not match pure single sources. The selectivity for time-slices and the ability to monitor multiple directions simultaneously is not an explicit prediction from our analyses, but selectivity appears to be an important ability to test in evaluating abilities of listeners in complex environments, with and without hearing-assist devices.

Turning to the ICC data, the distribution of ICC values within each environment can be illustrated by plotting the joint distribution of ICC values for 500-Hz time-slices and for 2000-Hz time-slices (based on envelopes, as described earlier). As an illustration, the 2000-Hz ICC is plotted versus the 500-Hz ICC for the combined data for all eight natural environments in Figure 8, as we did in Figure 5 for the SL. The data in this graph could be considered as exemplary for the coherence in bilateral stimuli that people encounter in daily life. It shows that the amount of TF slices that contain highly coherent information in both frequency bands is relatively small. The fraction above 0.95 in both frequencies

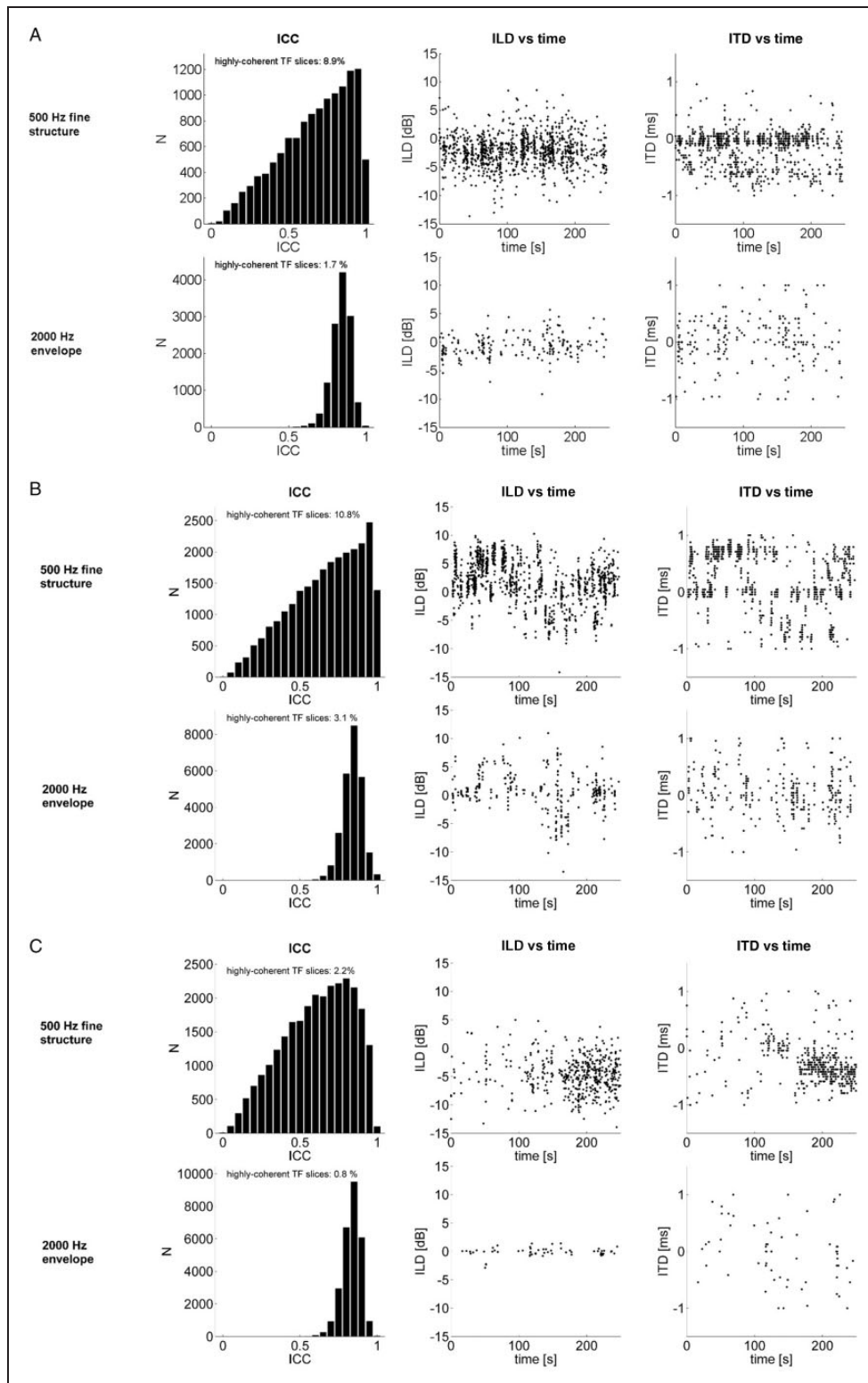


Figure 7. Interaural Parameters Analyzed in 20-ms Time-Slices for Several Natural Environments. A: An Inside environment (Restaurant). B: An Outside environment (City Walk). C: A Public Transport environment (In a Train). For each environment, the three panels in the upper row present data for the 500 Hz frequency band, and the lower row presents data for the 2000-Hz band. The distribution of ICC values are in the leftmost panel, with the fraction of time–frequency slices that are highly coherent given; the center panel gives the ILD values plotted versus time; and the right panel gives the ITD values plotted versus time. Only the highly correlated values ($ICC > 0.95$) are plotted in both cases. ICC = interaural cross-coherence; ITD = interaural time delay; ILD = interaural level difference.

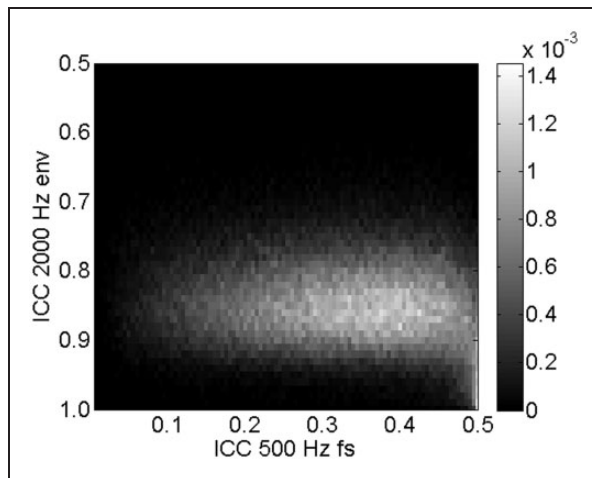


Figure 8. Combined coherence data for all eight natural environments in the 2000-Hz band (envelope) versus the 500-Hz band (fine-structure) in bins that are spaced by 0.01. Numbers are expressed as fractions of the total number of time-slices in the eight environments (143104). The data in this graph could be considered as exemplary for the coherence in binaural stimuli that people encounter in daily life. Note that the ICC values are scattered and that maximum values are about 0.0014. (For the virtual recording with female target at $+45^\circ$ and male masker at -45° , values are nearly all in the high coherent region with maximum values of 0.45.)

is less than 2%. Also, note that the distributions are wide, with a range from zero to unity for the 500-Hz case and with a range from roughly 0.6 to unity for the 2000-Hz envelopes. The higher values for the 2000-Hz case is expected because the envelopes are always positive, as noted earlier (cf. Bernstein & Trahiotis, 1996a, 1996b). For the anechoic virtual environment with female target at -45° and MALE masker at $\pm 45^\circ$ (Figure 6A), the fraction above an ICC of 0.95 in both frequencies is more than 20%. Note that this environment corresponds to configurations used in clinical testing. The general picture that arises is that stimuli in everyday life are much less coherent than stimuli used in clinical testing, though the exact numbers will depend on the specific choices that are made in selecting the specific environments. This implies that the number of instances in which one source is dominating both ears is relatively small, even with 20-ms intervals.

Discussion

Binaural sound recordings were made in a variety of natural environments. For comparison and reference purposes, virtual recordings were constructed with a female target, several types of maskers, and several spatial configurations. Recorded waveforms were

analyzed in terms of SL, based on modulation patterns, and in terms of interaural parameters, that is, distributions of ICC, ITD, and ILD estimated for 20-ms time-slices.

Speech-Likeness

Results for the SL measures in natural recordings (Figure 4) show large differences among the environments that are encountered in everyday life. Also, within the individual recordings, SL may vary substantially over different time intervals as well as show large differences between left and right ears. There is evidence that scenarios with adverse acoustic conditions occur in everyday life as do scenarios with highly intelligible speech. These results are generally in line with those reported by Smeds et al. (2015) and Wu et al. (2018). To make a better comparison with these studies, we crudely relate the SL measure for the virtual recordings to speech intelligibility, using results of speech recognition tests with the same stimuli conducted on normal-hearing listeners in a previous study (Smits et al., 2013). This indicative comparison is only done for the anechoic condition with a collocated target and LTASS masker (matching the circles in the upper left panel in Figure 2). For this case, the (virtual) target and masker are at the same location, directly in front of the listener, and the spectra of target and masker are the same. Therefore, the TMR values at the sources are the same as the TMR values at both ears. Smits et al. (2013) present speech recognition data as a function of SNR, which is the same as our TMR in this case, with speech and LTASS stimuli presented monaurally through headphones. In their study (Smits et al., 2013), 100% speech intelligibility is reached at an SNR (and TMR) of 0 dB. From Figure 2, upper left panel, it follows that a TMR of 0 dB (and 100% speech intelligibility) corresponds to a SL value of 8.2. Similarly, 50% speech intelligibility is reached at a TMR of -5 dB, which corresponds to a SL value of 4.7, and finally, 0% speech intelligibility is reached at a TMR of -10 dB, which corresponds to a SL value of 1.2. Note that clean speech (TMR ≥ 15 dB) corresponds to speech-intelligibility values above 12.2.

Using this crudely derived relationship between SL values and SNRs (SL values of 1.2, 4.7, 8.2, and 12.2, corresponding to SNRs of -10 , -5 , 0, and 15 dB, respectively), we can intuitively relate the results of this study to those of Smeds et al. and Wu et al. SNRs ranging from -10 to 30 dB and peaking at about an SNR of 8 dB in Wu's Figure 4 yield SL values ranging from 1.2 to more than 12.2 with a peak around 10. In the current data, SL in Inside environments shows a similar pattern (Figure 4A and B). In Outside environments (Figure 4C–E), lower values for SL are found;

furthermore, the distribution shows a second peak near zero, possibly reflecting instances with adverse conditions or conditions with no speech present at all. In the Public Transport environments (without intentional conversation), SL does not exceed 10 and only a peak toward low SNRs is found. When taken together, the distribution of SNR/SL in this study is broader than that of Wu et al. and more in line with the data reported by Smeds et al. (2015, Figure 4).

There are several differences in methodology that might account for the different results in different studies. We used an automated measure to characterize SL, without using subjective characterization by the subject. Hence, we presumably have included time-slices without speech at all, or time-slices that would not be considered as a communication situation. This might have led to a lower estimation of values for SL and consequently for the estimated SNR values. On the other hand, in the manual procedure, using characterization of time-slices, some relevant situations might be missed; furthermore, this manual method is complicated in relation to privacy issues.

Returning to Figure 5, in which the SL values for the right ear versus the left ear are plotted, combined over all natural environments, we consider our results with reference to speech-intelligibility results using the reference frame described just above by comparing SL values with speech-intelligibility values for the LTASS masking case. Note that these comparisons are only approximate because we do not have validated information about the relation between intelligibility and speech likeness for maskers other than LTASS noise. However, this way of plotting could be considered as exemplary for the SL that people encounter bilaterally in daily life, where “good speech-likeness” is defined as $SL \geq 8.2$ as corresponding to 100% speech intelligibility in the LTASS case. As an illustration, four subdivisions are distinguished in Figure 5 with solid black lines: (A) good SL for both ears, (B) good SL only on the right side, (C) good SL only on the left side, and (D) poor SL for both ears. In Subdivision A, using the indicative relation between SL and speech intelligibility, normal hearing listeners would have no difficulties in speech intelligibility and the exact spatial position would not be critical. In Subdivision B, listeners would achieve good speech intelligibility using their right ear but not their left, and in Subdivision C listeners would need their left ear. In Subdivision D, neither ear alone would achieve a good speech intelligibility score; however, if binaural coherence is sufficient, binaural unmasking could enhance the effective SNR up to 5 dB (e.g., Goverts & Houtgast, 2010; Levitt & Rabiner, 1967a, 1967b). This would effectively improve performance in Subdivision D.

For an individual listener with impaired hearing, the same four subdivisions of functional abilities can be indicated, but the functional boundaries would be at higher values of SL. Example boundaries are also included in Figure 5 with gray lines. Hearing impairment can, even while wearing hearing aids, lead to an increase in the TMR in LTASS noise required for 50% intelligibility and a shallowing of the psychometric curve that varies over listeners. We describe an exemplary shift of around 3 dB (-2 dB vs. -5 dB). As a consequence, the TMR for 100% intelligibility shifts by about 4 to 5 dB in this example corresponding to a SL of about 11. This is illustrated by the solid gray lines. These changes imply that individuals with sensory-neural hearing impairment will encounter *more* time-slices in Subdivision D; the number of slices in the other subdivisions will change correspondingly.

Binaural unmasking can be functional in individuals with impaired hearing when they have adequate bilateral hearing-aid fitting (e.g., Goverts & Houtgast, 2010); however, in many cases it will be reduced, even if the binaural coherence in the environment is substantial. So, additional signal processing schemes, directional microphones, and remote listening solutions would benefit these individuals encountering sound stimuli in Subdivision D. From Figure 4C to E, it also becomes clear that, for individuals with bilateral hearing loss who are rehabilitated unilaterally, the fraction of time-slices that show significant differences in SL between the ears, which is presumably related to the number of situations with useful speech on one side and nonspeech interference on the other, can be substantially reduced if a second device is used. To conclude, these exploratory data suggest that the better-ear effect is functionally important, which provides support for the prescription of bilateral instead of unilateral amplification where appropriate. It should also be noted that there are large differences in SL *within and between* the situations that individuals with hearing impairment encounter.

The results also show many cases with large differences in SL between the left and right waveforms, with the better-ear changing in different time intervals, suggesting that it is important to have both ears receiving stimuli, especially in conditions that do not obviously favor a single ear. This suggestion results from the spatial configurations and movements of sources as well as the short-time variation of which source in the environment is dominant. Figure 5 presents an overview of interaural differences in SL for the eight environments combined. It can be shown that the overall median value is about unity (range 0–13). Smeds et al. (2015, Table 2) report median differences in SNR between the better and worse ear for different environments. The overall median value is 3 dB (range 0–8 dB). The different sizes of interaural differences in SL are probably caused by the smaller

time frames used in this study. Note also that Smeds et al. find the largest interaural difference for the Public Transport, whereas in this study, the largest values are found in Outside environments (Figure 4C–E, second column), presumably reflecting the fact that Smeds et al. restricted their analysis to time frames with intentional communication. Taken together, these factors highlight the importance of both the ability to listen to both ears (sometimes individually and sometimes together) and the ability to select the ear or mode of binaural processing within short-time intervals, all as part of the optimization of binaural listening.

Modeling studies of speech intelligibility, including spatially distributed maskers of various types, show that individuals with normal hearing can use binaural processing for either selecting optimal TF units (cf. Mi et al., 2017) or suppressing the masker (cf. Beutelman & Brand, 2006; Beutelmann et al., 2010; Lavandier & Culling, 2010; Wan et al., 2010, 2014). For an individual with hearing impairment, the short-time boundaries between using one of the two ear signals or using binaural processing are determined by speech recognition capacities, that is, what amount of speech information (corresponding to a specific SL value) is needed for intelligibility, and what processing is available. Little is known about these abilities.

Interaural Differences

Results from the interaural difference measures show that, in the 500-Hz frequency band, the 20-ms time-slice is short enough to capture the times that single sources dominate the binaural inputs when there are multiple speech sources in the environment. Importantly, results show that the natural recordings contain a relatively small amount of coherent time-slices (only 0.5–15.5% of time-slices had an interaural coherence above 0.95) when compared with the virtual recordings (44% and 35% for the T-45M45 configuration with MALE and LTASS masker, respectively). Only the reverberant virtual case showed a comparably low percentage (5%). These coherent time-slices contain binaural cues (ITD and ILD) that can be used to select intervals with good speech intelligibility for individual sources. Within these coherent time-slices, the interaural differences observed as a function of time show good (subjective) correspondence with the subjective impressions of dominant sources when listening to the natural recordings. Also, the nature of the patterns of interaural differences as a function of time vary over the environments, as one would expect.

The distributions of interaural parameters, as seen for several environments in Figure 7, illustrate that 20-ms time-slices can be used to estimate the interaural parameters of individual sources. It follows that this 20-ms

interval, which has been used in binary masking paradigms (e.g., Best et al., 2017), is short enough to allow individual speech sources to dominate in many time intervals. By listening to the recordings while observing the distributions of interaural differences over time, we conclude that these interaural differences could be used to identify the dominant source and possibly to TF filter the stimulus to provide useful information about a selected source. In general, the experience of listening in combination with the analysis of the ITDs for different time-slice lengths suggest that processing of information in time intervals on the order of 20 ms provides useful information for separating sources. It is also notable that the ILD distributions do not show distinct peaks, even for cases when distinct peaks are seen in the ITD distributions (i.e., for time-slices with high ICC values).

The ability of listeners to make use of rapid changes in interaural parameters is an unresolved question. We used TF slices of 20-ms duration to provide information at a high-resolution time scale so that we would be able to see the variations, and we chose 20-ms to be consistent with other high-resolution studies (Best et al., 2017; Culling et al., 2006). The ability of the auditory processing to make use of such rapid changes in interaural parameters is not yet clear. The concept of “binaural sluggishness” is needed to understand a variety of binaural psychophysical data, and it is incorporated into many binaural models, particularly in the area of binaural detection. These models typically assume a temporal smoothing to generate available binaural decision variables with a time constant on the order of 100 ms (e.g., Durlach, 1963; Hauth & Brand, 2018); however, the conditions under which sluggishness has an impact have not been determined. We believe that more experiments in this general area will provide useful data.

Clinical Relevance

This study indicates that the acoustics of everyday life situations vary substantially, and many situations contain sparse information for speech recognition. Thus, selective listening to temporal slices is important in complex environments. The normal auditory-cognitive system can function adequately in the majority of environments using these cues; however, in cases of auditory impairment and/or reduced top-down resources, everyday life is a challenging condition.

The substantial fraction of time-slices with a single good ear as measured by SL suggests that the better-ear effect is relevant for daily functioning, which supports the prescription of bilateral amplification for appropriate candidates. This also supports the notion that patients miss substantial information if they do

not have the use of two ears (e.g., because of unilateral deafness).

In addition, for individuals with impaired hearing, our analyses suggest that everyday environments contain relatively high numbers of time-slices in which neither ear receives adequate SL. This situation may improve with bilateral amplification enabling binaural unmasking, but in many cases advanced signal-processing schemes or remote microphones would be needed to enhance the SNR (and hence the SL) to achieve good intelligibility.

Finally, restoring/rehabilitating binaural function is important. This study shows that, even in situations with low interaural coherence, there is relevant binaural information in realistic stimuli for the localization and separation of sources. This information should be processed and transferred optimally to the auditory system. Listeners with impaired hearing should be trained to optimally make use of these cues, and hearing-assist devices should be designed to maintain these important cues.

Assessing Speech Recognition in Clinical Testing

Results show that there are large differences in SL within and between the situations that individuals with hearing impairments encounter. Furthermore, it has become clear that interaural coherence is low and variable in realistic conditions. In clinical consultations and assessment, this variability should be taken into account. Clinical tests typically assess speech recognition capacity in a well-defined setting, usually highly coherent stimuli, like the virtual recording with female target speech and male masker in this study. These tests usually target the condition of just-intelligible speech, generally at negative TMRs (i.e., low values for SL). This study shows that such tests are only partly representative for daily life and that there is broad scope for ecologically valid tests. If new tests are to move toward higher ecological validity, they should include conditions with higher and more variable SL values. There have been some attempts to add realism to speech tests, for example, Best et al. (2015) and Culling (2016). Overall, improved tests should not necessarily mimic realistic configurations, but merely include more temporal dynamics in SL values and in interaural parameters and also less coherent stimuli. Furthermore, Weisser and Buchholz (2019) suggest that speech and noise should not be considered independently. SNR should follow operational ranges that occur in realistic acoustic conditions as were investigated in their paper. Furthermore, with regard to spatial configuration, clinical testing of speech recognition in spatially separated configurations occurs typically in T0M+45 and/or T-45M+45 configurations. Most of the recorded natural time-slices (except for the “Walk”

recording) have interaural differences in SL that fall within the range that is found in virtual T0M+45 or T-45M+45 recordings. This suggests that current speech tests with T-45M+45 configurations are ecologically consistent with natural environments in terms of interaural differences.

Future Research

Research studies should be designed to specifically investigate the mechanisms underlying the apparent capacities of human beings to process the sparse distribution of coherent TF slices and to make optimal selections of information. One possible direction is performing psychophysical investigations of speech intelligibility using the realistic recordings as a masker as was done by Weisser and Buchholz (2019).

Directions for future research include the development of clinically useful tests that can be used to evaluate performance in multisource environments and that can be easily applied as part of the hearing-aid fitting and evaluation process. These tests would naturally include multiple sources with fluctuations of short-term power that are comparable to fluctuations in speech waveforms. Another area of research that may be helpful would be to make recordings in environments for specific populations (e.g., children, older adults, and hearing impaired), who may encounter special problems in specific environments (e.g., classrooms, gyms, large dining rooms, group meetings, and parties with multiple small groups having conversations). We may also benefit from recordings with longer durations that would allow estimation of relative occurrence of special situations within their acoustic environments.

Finally, recordings like those of this study should be used to evaluate signal processing in hearing-assist devices to investigate what output is provided to the eardrum (or to the auditory nerve) in response to the realistic stimuli and then specifically evaluate how available signal cues are preserved in binaural and bi-modal (electric and acoustic) processing.

Conclusions

Overall, the SL results from realistic environments show large variability within and between environments and also show many time-slices with a single good ear. Furthermore, the interaural parameter results show that the natural recordings contain a relatively small proportion of time-slices with high coherence compared with the virtual recordings. In other words, the information available to the binaural hearing system in realistic scenarios is sparse. The normal auditory-cognitive system can function adequately in the majority of environments; however, in cases of auditory impairment

and/or reduced top-down resources, everyday life has many challenging acoustic conditions.

For individuals with impaired hearing, our analyses suggest that everyday environments contain relatively high numbers of time-slices in which neither ear receives adequate speech information. This situation may improve with bilateral amplification, but in many cases advanced signal-processing schemes or remote microphones would be needed to enhance the SNR (and hence the SL) to achieve good intelligibility.

Furthermore, restoring/rehabilitating binaural function is important. This study shows that, even in situations with low interaural coherence, there is relevant binaural information in realistic stimuli for the localization and separation of sources. This information should be processed and transferred optimally to the auditory system. Listeners with impaired hearing should be trained to optimally make use of these cues, and hearing-assist devices should be designed to maintain these cues.

Finally, there are implications for improving clinical tests to assess speech recognition capacity. Current tests typically use well-defined setting with highly coherent stimuli, like the virtual recording with female target speech and male masker in this study. These tests usually target the condition of just-intelligible speech at negative TMR. This study shows that such tests are only partly representative for daily life and that there is broad scope for improved, ecologically valid tests. Such tests should not necessarily mimic realistic configurations but should at least include more temporal dynamics in SL values and in interaural parameters and also less coherent stimuli in order to enhance ecological validity.

Acknowledgments

We thank Andrew Brughera and Hans van Beek for helping with hardware and software issues, Alieke Breure and Krista Jansen for making recordings and contributing to the study. We also thank Virginia Best and Tammo Houtgast for useful comments on an earlier version of this manuscript. Finally, we acknowledge the many constructive suggestions of the three reviewers and our editor, all four of whom helped us to evaluate and explain our results.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Much of this work was supported by the Hearing Industry Research Consortium. They supported the primary recordings, simulations, and analyses presented here. In addition, work

was supported by the National Institutes of Health (NIDCD Grant DC000100).

ORCID iD

S. Theo Goverts  <https://orcid.org/0000-0002-6887-5909>

References

- Bernstein, L. R., & Trahiotis, C. (1996a). On the use of the normalized correlation as an index of interaural envelope correlation. *Journal of the Acoustical Society of America*, *100*, 1754–1763. <https://doi.org/10.1121/1.416072>
- Bernstein, L. R., & Trahiotis, C. (1996b). The normalized correlation: Accounting for binaural detection across center frequency. *Journal of the Acoustical Society of America*, *100*, 3774–3784. <https://doi.org/10.1121/1.417237>
- Best, V., Keidser, G., Buchholz, J., & Freeston, K. (2015). An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment. *International Journal of Audiology*, *54*(10), 682–690. <https://doi.org/10.3109/14992027.2015.1028656>
- Best, V., Mason, C.R., Swaminathan, J., Roverud, E., & Kidd, G. (2017). Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures. *Journal of the Acoustical Society of America*, *141*(1), 81–91. <https://doi.org/10.1121/1.4973620>
- Beutelmann, R., and Brand, T. (2006) Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, *120*(1), 331–42. Doi: 10.1121/1.2202888.
- Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *Journal of the Acoustical Society of America*, *127*(4), 2479–2497. <https://doi.org/10.1121/1.3295575>
- Bregman, A. S. (1990). *Auditory scene analysis*. The MIT Press
- Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple talker conditions. *Acta Acustica united with Acustica*, *86*, 117–128. DOI: DOI 10.3758/s13414-015-0882-9
- Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of the Acoustical Society of America*, *120*(6) 4007–4018. <https://doi.org/10.1121/1/2363929>
- Cosentino, S., Marquardt, T., McAlpine, D., Culling, J., & Falk, T. H. (2014). A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals. *Journal of the Acoustical Society of America*, *135*, 796–807. <https://doi.org/10.1121/1.4861239>
- Culling, J. F. (2016). Speech intelligibility in virtual restaurants. *Journal of the Acoustical Society of America*, *140*(4), 2418–2426. <https://doi.org/10.1121/1.4964401>
- Culling, J.F., Edmonds, B.A., Hodder, K.I. (2006) Speech perception from monaural and binaural information. *Journal of the Acoustical Society of America*, *119*(1), 559–65. Doi: 10.1121/1.2140806

- Culling, J. F., Hawley, M. L., & Litovsky, R. Y. (2004). The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *Journal of the Acoustical Society of the America*, *116*, 1057–1065. <https://doi.org/10.1121/1.1772396>
- Culling, J. F., Jelfs, S., Talbert, A., Grange, J., & Backhouse, S. S. (2012). The benefit of bilateral versus unilateral cochlear implantation to speech intelligibility in noise. *Ear and Hearing*, *33*(6), 673–683. <https://doi.org/10.1097/AUD.0b013e3182587356>
- Dubbelboer, F., & Houtgast, T. (2008). The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *Journal of the Acoustical Society of the America*, *124*(6), 3937–3946. <https://doi.org/10.1121/1.3001713>
- Durlach, N.I. (1963) Equalization and cancellation theory of binaural masking-level differences. *Journal of the Acoustical Society of America*, *35*(8), 1206–1218. doi: 10.1121/1.1918675
- Falk, T. H., Zheng, C., & Chan, W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*, 1766–1773. <https://doi.org/10.1109/TASL.2010.2052247>
- Festen, J.M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of the America*, *88*(4), 1725–1736. <https://doi.org/10.1121/1.400247>
- Goverts, S.T., & Houtgast, T. (2010). The binaural intelligibility level difference in hearing-impaired listeners: The role of supra-threshold deficits. *Journal of the Acoustical Society of the America*, *127*, 3073–3084. <https://doi.org/10.1121/1.3372716>
- Hauth, C. F., & Brand, T. (2018). Modeling sluggishness in binaural unmasking of speech for maskers with time-varying interaural phase differences. *Trends in Hearing*, *22*, 1–10. <https://doi.org/10.1177/2331216517753547>
- Hawley, M. L., Litovsky, R. Y., & Culling, J. S. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of the America*, *115*(2), 833–843. <https://doi.org/10.1121/1.1639908>
- Houtgast, T., & Steeneken, H. J. M. (1972). *Envelope spectrum and intelligibility of speech in enclosures* [Conference session]. In Proceedings of IEEE Speech Conference, Newton, MA, United States. pp. 392–395.
- Jørgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *Journal of Acoustical Society of America*, *130*(3), 1475–1487. <https://doi.org/10.1121/1.3621502>
- Jørgensen, S., Ewert, S. D., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *Journal of Acoustical Society of America*, *134*, 436–446. <https://doi.org/10.1121/1.4807563>
- Kidd, G., Jr., Mason, C. R., Best, V., Roverud, E., Swaminathan, J., Jennings, T., Clayton, K., & Colburn, H. S. (2019). Determining the energetic and informational components of speech-on-speech masking in listeners with sensorineural hearing loss. *Journal of Acoustical Society of the America*, *145*(1), 440–457. <https://doi.org/10.1121/1.5087555>
- Lavandier, M., & Culling, J. F. (2010). Prediction of binaural speech intelligibility against noise in rooms. *Journal of the Acoustical Society of the America*, *127*, 387–399. <https://doi.org/10.1121/1.3268612>
- Levitt, H., & Rabiner, L. R. (1967a). Binaural release from masking for speech and gain in intelligibility. *Journal of the Acoustical Society of the America*, *42*, 601–608. <https://doi.org/10.1121/1.1910629>
- Levitt, H., & Rabiner, L. R. (1967b). Predicting binaural gain intelligibility and release from masking for speech. *Journal of the Acoustical Society of the America*, *42*, 820–829. <https://doi.org/10.1121/1.1910654>
- Marrone, N., Mason, C. R., & Kidd, G., Jr. (2008a). Evaluating the benefit of hearing aids in solving the cocktail party problem. *Trends in Amplification*, *12*, 300. <https://doi.org/10.1177/1084713808325880>
- Marrone, N., Mason, C. R., & Kidd, G., Jr. (2008b). Tuning in the spatial dimension: Evidence from a masked speech identification task. *Journal of Acoustical Society of the America*, *124*(2), 1146–1158. <https://doi.org/10.1121/1.2945710>
- Marrone, N., Mason, C. R., & Kidd, G., Jr. (2008c). The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *Journal of Acoustical Society of the America*, *124*(5), 3064–3075. <https://doi.org/10.1121/1.2980441>
- Mi, J., Groll, M., & Colburn, H.S. (2017). Comparison of a target-equalization-cancellation approach to a localization approach to source separation. *Journal of Acoustical Society of the America*, *142*(5), 2933–2941. <https://doi.org/10.1121/1.5009763>
- Młynarski, W., & Jost, J. (2015). Statistics of natural binaural sounds. *PLoS One*, *9*(10), e108968. <https://doi.org/10.1371/journal.pone.0108968>
- Sabine, W. C. (1900/1915). *Collected papers on Acoustics (Peninsula, Los Altos, CA)*. Google Scholar.
- Shinn-Cunningham, B. G., Desloge, J. G., & Kopco, N. (2001). *Empirical and modeled acoustic transfer functions in a simple room: Effects of distance and direction*. In Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 19–24 October 2001, 183–186.
- Smeds, K., Wolters, F., & Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of the American Academy of Audiology*, *26*(2), 183–196. <https://doi.org/10.3766/jaaa.26.2.7>
- Smits, C., Goverts, S. T., & Festen, J. M. (2013). The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *Journal of the Acoustical Society of the America*, *133*, 1693–1706. <https://doi.org/10.1121/1.4789933>
- Versfeld, N. J., Daalder, L., Festen, J. M., & Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *Journal of the Acoustical Society of the America*, *106*, 1671–1684.
- Wagener, K. C., Hansen, M., & Ludvigsen C. (2008). Recording and classification of the acoustic environment

- of hearing aid users. *Journal of the American Academy of Audiology*, 19(4), 348–370. <https://doi.org/10.3766/jaaa.19.4.7>
- Wan, R., Durlach, N. I., & Colburn, H. S. (2010). Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. *Journal of the Acoustic Society of the America*, 128, 3678–3690.
- Wan, R., Durlach, N. I., & Colburn, H. S. (2014). Application of a short-time version of the equalization cancellation model to speech intelligibility experiments with speech maskers. *Journal of the Acoustic Society of the America*, 136, 768–776. <https://doi.org/10.1121/1.4884767>
- Weisser, A., and Buchholz, J. (2019) Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions. *Journal of the Acoustic Society of the America*, 145, 349–360. DOI: 10.1121/1.5087567
- Wu, Y. H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., & Oleson, J. (2018). Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss. *Ear Hearing*, 39(2), 293–304. <https://doi.org/10.1097/AUD.0000000000000486>