Effect of Channel Interaction on Vocal Cue Perception in Cochlear Implant Users

Waldo Nogueira, Nawal El Boghdady, Florian Langner, Etienne Gaudrain, Deniz Baskent

To cite this version:

Waldo Nogueira, Nawal El Boghdady, Florian Langner, Etienne Gaudrain, Deniz Baskent. Effect of Channel Interaction on Vocal Cue Perception in Cochlear Implant Users. 2020. hal-03015084
Effect of Channel Interaction on Vocal Cue Perception in Cochlear Implant Users

Waldo Nogueira\textsuperscript{1,◊}, Nawal El Boghdady\textsuperscript{2,3◊,*}, Florian Langner\textsuperscript{1}, Etienne Gaudrain\textsuperscript{4,2,3}, Deniz Başkent\textsuperscript{2,3},

1. Department of Otolaryngology, Medical University Hannover and Cluster of Excellence Hearing4all, Hanover, Germany
2. University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology, Groningen, Netherlands
3. University of Groningen, Research School of Behavioral and Cognitive Neurosciences, Groningen, Netherlands
4. Lyon Neuroscience Research Center, CNRS UMR 5292, INSERM U1028, University Lyon 1, Lyon, France
◊. Shared first authorship
*. Corresponding author: n.el.boghdady@umcg.nl
ABSTRACT

Speech intelligibility in multi-talker settings is challenging for most CI users. One possibility for this limitation is the suboptimal representation of vocal cues in implant processing, such as the fundamental frequency (F0), and the vocal tract length (VTL). Previous studies suggested that while F0 perception depends on both spectral and temporal cues, VTL perception relies largely on spectral cues. To investigate how spectral smearing in CIs affects vocal cue perception in speech-on-speech (SoS) settings, parallel channels were simultaneously stimulated in 14 Advanced Bionics users to simulate channel interaction. Three such patterns were created: Sequential (stimulation of 2 simultaneous electrodes), Paired (4 electrodes), and Triplet stimulation (6 electrodes). F0 and VTL just-noticeable differences (JNDs; task 1), in addition to SoS intelligibility (task 2) and comprehension (task 3) were measured for each stimulation strategy. In tasks 2 and 3, four maskers were used: the same female talker, a male voice obtained by manipulating both F0 and VTL (F0+VTL) of the original female speaker, a voice where only F0 was manipulated, and a voice where only VTL was manipulated. JNDs were measured relative to the original voice for the F0, VTL, and F0+VTL manipulations. When spectral smearing was increased, a significant deterioration in performance was observed for all tasks, with no differences between Sequential and Paired stimulation. These results imply that CI users may tolerate certain amounts of channel interaction without significant reduction in performance on tasks relying on voice perception. This points to possibilities for utilizing parallel stimulation in CIs for reducing power consumption.

Keywords: channel interaction; cochlear implant; speech-on-speech; F0; VTL

INTRODUCTION

Cochlear implants (CIs) are devices that can restore hearing in people suffering from profound hearing loss. Although many CI users obtain good speech performance in quiet, their speech intelligibility drops significantly in the presence of two or more simultaneous speakers (e.g.,
The performance of CI users in such a speech-on-speech (SoS) scenario has been shown in a previous study (El Boghdady, Gaudrain, & Başkent, 2019) to be correlated with their sensitivity to two important voice cues defining the voices of the target and masker speakers: the fundamental frequency (F0) and the vocal tract length (VTL) of the speaker.

The speaker’s F0 induces the percept of the voice pitch and is usually lower for adult males than adult females (Peterson & Barney, 1952; Smith & Patterson, 2005). These F0 cues are usually encoded in both the temporal envelope and the cochlear location of excitation (e.g., Carlyon & Shackleton, 1994; Licklider, 1954; Oxenham, 2008), which gives these cues a spectrotemporal nature.

The VTL correlates with the speaker’s physical (Fitch & Giedd, 1999) and perceived height (Ives, Smith, & Patterson, 2005; Smith, Patterson, Turner, Kawahara, & Irino, 2005), and is usually longer for adult males than for adult females. VTL cues are usually represented through the speech spectral envelope (Chiba & Kajiyama, 1941; Fant, 1960; Lieberman & Blumstein, 1988; Müller, 1848; Stevens & House, 1955). Shortening VTL results in the stretching of the spectral envelope towards higher frequencies on a linear frequency scale, while elongating VTL results in the compression of the spectral envelope towards lower frequencies. This means that VTL cues can be largely encoded in the relationship between the peaks in the spectral envelope of the signal. Hence, the adequate representation of both F0 and VTL cues would be expected to require sufficient spectrotemporal resolution.

Information transmitted by the CI is usually spectrotemporally degraded (Fu & Nogaki, 2005; Fu, Shannon, & Wang, 1998; Henry & Turner, 2003; Nelson & Jin, 2004; Winn, Won, & Moon, 2016). Spectrotemporal resolution in the implant depends on a number of factors, such as the amount of channel interaction between adjacent electrodes and the subsequent effective number of spectral channels (for a review, see Başkent, Gaudrain, Tamati, & Wagner, 2016). Because
of the conductive fluid filling the cochlea, current spreads between neighboring electrodes resulting in channel interaction (e.g. Boëx, de Balthasar, Kós, & Pelizzone, 2003; De Balthasar et al., 2003; Hanekom & Shannon, 1998; Shannon, 1983; Townshend & White, 1987), and the subsequent reduction in the effective number of spectral channels. The literature has demonstrated that CI listeners do not usually have access to more than 8 effective spectral channels (Friesen, Shannon, Başkent, & Wang, 2001; Qin & Oxenham, 2003) and that significant channel interaction not only impairs speech and phoneme perception (e.g. Friesen et al., 2001; Fu & Shannon, 2002; Qin & Oxenham, 2003), but also voice cues as well (Gaudrain & Başkent, 2015). In the latter study, Gaudrain and Başkent (2015), using vocoder simulations of CI processing, have demonstrated that as channel interaction increases (simulated as the shallowness of the vocoder filter slopes), the sensitivity to VTL cues deteriorates. Thus, the poor spectrotemporal resolution in CIs is also expected to influence voice differences between target and masker speakers in SoS scenarios.

This study aims at assessing the effects of such channel interaction (and resulting spectral resolution) on SoS and voice cue perception in CI listeners by using simultaneous stimulation of different channels. Beyond the purpose of understanding how crucial spectrotemporal cues are for F0, VTL, and SoS perception, there is also a potential benefit in using parallel stimulation, since it was originally proposed in the literature as a method of reducing power consumption (e.g., Büchner, Frohne, Battmer, & Lenarz, 2005; Frijns, Kalkman, Vanpoucke, Bongers, & Briaire, 2009; Langner, Saoji, Büchner, & Nogueira, 2017). One way of achieving this is to decrease the maximum stimulation current required to stimulate the auditory nerve. For instance, by stimulating two adjacent electrodes in the cochlea it is possible to reduce the amount of current by half to achieve the same loudness percept as that from single electrode stimulation, since the current is distributed between both electrodes. Additionally, it is possible to introduce simultaneously stimulated parallel channels, for example Paired (two pairs of simultaneous channels, with one channel consisting of 2 adjacent electrodes simultaneously
stimulated at a time) and Triplet stimulation (three pairs of simultaneous channels), to reduce
the maximum current delivered by the implant by a factor of 17% and 44%, respectively. With
Paired stimulation, it is possible to double the pulse duration with respect to Sequential
stimulation without changing the stimulation rate of the implant. In terms of performance,
Langner et al. (2017) showed no degradation of speech performance under stationary
background noise for Paired stimulation. However, the same study also showed that increasing
the number of parallel channels to three, such as in Triplet stimulation, causes a significant drop
in speech intelligibility in comparison to Sequential stimulation (a single pair of simultaneous
channels). From these results, it was suggested that Paired stimulation may be a good candidate
for reducing power consumption in CI users, however, more detailed speech performance
measures are required to assess the potential effects of adding parallel channels (spectral
smearing and channel interaction) on speech intelligibility. Thus, another goal of this study, if
only degradations were to be observed, was to determine the level of parallel channel
stimulation that could be acceptable for voice cue and SoS perception, without significant
reduction in performance.

Three research questions were addressed in this study: 1) whether increasing the number of
parallel stimulated channels (increasing channel interaction) decreases the sensitivity to F0 and
VTL differences in CI users (task 1), and 2) whether this effect is also reflected as a reduction
in SoS perception (tasks 2 and 3). The expectations were that these effects should be larger for
VTL compared to F0 perception, because VTL is a primarily spectral cue, while F0 cues could
still be preserved in the temporal aspect of the signal even if the spectral component is
compromised. The third research question was 3) whether some parallel channel stimulation
could be deployed for reducing power consumption without significantly impairing voice cue
and SoS perception.
I. METHODS

A. Participants

Twelve native German CI users with Advanced Bionics (AB) devices were recruited from the clinical database of the Medizinische Hochschule Hannover (MHH) based on their clinical speech intelligibility scores in quiet and in noise. To ensure that participants could perform the SoS tasks, the inclusion criteria were to have a speech intelligibility score higher than 80% in quiet and 20% in noise at a +10 dB signal-to-noise ratio on the Hochmair-Schulz-Moser (HSM) sentence test (Hochmair-Desoyer, Schulz, Moser, & Schmidt, 1997). Table 1 shows the demographics of the CI users. Only 8 (P05-P12) of the 12 participants participated in the SoS comprehension task.

Table 1. Demographics for CI users recruited. All durations in years are calculated based on the date of testing. Progressive hearing loss refers to participants who experienced minimal hearing loss that gradually progressed until they fulfilled the criteria for acquiring a CI.

<table>
<thead>
<tr>
<th>Participant number</th>
<th>Gender</th>
<th>Age at testing (y)</th>
<th>Implant</th>
<th>Duration of device use (y)</th>
<th>Duration of hearing loss (y)</th>
<th>Etiology</th>
<th>Clinical speech-in-quiet scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>P01</td>
<td>M</td>
<td>20</td>
<td>Helix</td>
<td>4.0</td>
<td>0</td>
<td>Unknown</td>
<td>100%</td>
</tr>
<tr>
<td>P02</td>
<td>F</td>
<td>48</td>
<td>Helix</td>
<td>8.7</td>
<td>0.61</td>
<td>Acute</td>
<td>100%</td>
</tr>
<tr>
<td>P03</td>
<td>M</td>
<td>55</td>
<td>Mid-Scala</td>
<td>3.8</td>
<td>Progressive</td>
<td>Unknown</td>
<td>96%</td>
</tr>
<tr>
<td>P04</td>
<td>M</td>
<td>58</td>
<td>Mid-Scala</td>
<td>2.5</td>
<td>Progressive</td>
<td>Unknown</td>
<td>100%</td>
</tr>
<tr>
<td>P05</td>
<td>M</td>
<td>47</td>
<td>Mid-Scala</td>
<td>5.5</td>
<td>1.5</td>
<td>Acute</td>
<td>100%</td>
</tr>
<tr>
<td>P06</td>
<td>M</td>
<td>43</td>
<td>Helix</td>
<td>10.5</td>
<td>Progressive</td>
<td>Acute</td>
<td>98.11%</td>
</tr>
<tr>
<td>P07</td>
<td>F</td>
<td>51</td>
<td>Helix</td>
<td>11.4</td>
<td>0</td>
<td>Genetic</td>
<td>90.56%</td>
</tr>
<tr>
<td>P08</td>
<td>F</td>
<td>70</td>
<td>Helix</td>
<td>2.6</td>
<td>5.24</td>
<td>Unknown</td>
<td>100%</td>
</tr>
<tr>
<td>P09</td>
<td>M</td>
<td>51</td>
<td>Mid-Scala</td>
<td>5.6</td>
<td>Progressive</td>
<td>Unknown</td>
<td>95.25%</td>
</tr>
<tr>
<td>P10</td>
<td>F</td>
<td>46</td>
<td>Helix</td>
<td>9.6</td>
<td>Progressive</td>
<td>Acute</td>
<td>100%</td>
</tr>
<tr>
<td>P11</td>
<td>F</td>
<td>49</td>
<td>Helix</td>
<td>8.2</td>
<td>0.05</td>
<td>Acute</td>
<td>70.75%</td>
</tr>
<tr>
<td>P12</td>
<td>M</td>
<td>65</td>
<td>Helix</td>
<td>10</td>
<td>Progressive</td>
<td>Unknown</td>
<td>99.06%</td>
</tr>
</tbody>
</table>
B. Voice cue manipulations

F0 and VTL cues were manipulated relative to those of the original speaker of the corpus in each experiment using the *Speech Transformation and Representation based on Adaptive Interpolation of weighted spectrogram* (STRAIGHT; Kawahara & Irino, 2005). Increasing/decreasing F0 in STRAIGHT is implemented by shifting the pitch contour of the original speech upwards/downwards by a number of semitones (12\textsuperscript{th} of an octave; st) towards higher/lower frequencies relative to the average F0 of the stimulus. Shortening/elongating VTL is implemented by expanding/compressing the spectral envelope of the signal towards higher/lower frequencies.

Figure 1. [ΔF0, ΔVTL] plane, with the reference female speaker from experiment 2 shown as the solid black circle at the origin of the plane. Decreasing F0 and elongating VTL yields deeper-sounding male-like voices, while increasing F0 and shortening VTL yields child-like voices. The dashed ellipses are based on the data from Peterson and Barney (1952), which were normalized to the reference female speaker, and indicate the ranges of typical F0 and VTL differences between the reference female speaker and 99% of the population. The red crosses indicate the voice vectors from the origin of the plane along which the JNDs were measured in experiment 1, and the 4 different combinations of ΔF0 and ΔVTL used in both experiments 2 and 3.
Figure 1 shows the F0 and VTL values (red crosses) used in the current studies on the [ΔF0, ΔVTL] plane. The red crosses indicate the voice vectors (directions) from the origin of the plane along which the JNDs were measured in experiment 3 (along negative ΔF0, along positive ΔVTL, and along the diagonal passing through ΔF0 = -12 st, and ΔVTL = +3.8 st). In addition, they represent the 4 combinations of F0 and VTL differences between the masker and target speakers in experiments 2 and 3. The solid black circle at the origin on the plane indicates the voice of the original female speaker from the corpus used in experiment 2. The dashed ellipses encompass the range of relative F0 and VTL differences between the original female speaker 99% of the population as calculated from the Peterson and Barney study (1952). This calculation was performed by normalizing the data provided by Peterson and Barney relative to the voice parameters of the original female speaker of the corpus, who had an average F0 of about 218 Hz and an estimated VTL of around 13.97 cm. The original female speaker’s VTL was estimated using the method of Ives et al. (2005) and the data from Fitch and Giedd (1999), assuming an average height of about 166 cm for the speaker based on growth curves for the German population (Bonthuis et al., 2012; Schaffrath Rosario, Schienkiewitz, & Neuhauser, 2011). ΔVTL is oriented upside down to indicate that positive ΔVTLs yield a decrease in the frequency components of the spectral envelope of the signal.
Figure 2. Spectrograms of two German tokens (/da/ (top row) and /ɡɔ/ (bottom row)) shown for each voice.  

First column from left: original female speaker from the corpus; second column from left: effect of decreasing F0 by 12 st on the spectrogram; third column from left: effect of elongating VTL by 3.8 st; fourth column from left: effect of both decreasing F0 by 12 st and elongating VTL by 3.8 st relative to the voice of the original female speaker.

Figure 2 shows the effect of manipulating F0 and VTL on the spectrograms of two German tokens. The rows represent the different tokens, while the column represent the voice manipulation [no manipulation (original female speaker), F0, VTL, or both F0 and VTL]. Notice that as F0 decreases, the number of glottal pulses also decreases, and as VTL is elongated, the spectral content of the signal is compressed towards lower frequencies along a linear frequency scale. In addition, decreasing F0 and elongating VTL together yield less glottal pulses which are also compressed towards lower frequencies.
C. F120 Sound Coding Strategies (Sequential, Paired and Triplet)

1. Fidelity F120 Sound Coding Strategy

The Fidelity 120 (F120) in Advanced Bionics devices is a sound coding strategy that processes the audio signal through an automatic gain control. Next, a spectral analysis is performed using a short time fast Fourier transform (STFFT) to compute the slow varying envelopes in each analysis band. In parallel, the spectrum is analyzed in more detail using a spectral peak locator to estimate the most dominant frequency component in each analysis band. Finally, the slowly varying envelopes are logarithmically compressed into the electric dynamic range of each participant between the threshold and the most comfortable level. Each analysis band is assigned to two simultaneously stimulated electrodes [Figure 3 (B)]. The current ratio between these two electrodes is derived from the spectral peak locator forming a current steered – or virtual – channel. For a given analysis band \(k\), a pair of electrodes are simultaneously stimulated, one with current \(I_k \cdot \alpha\) and the adjacent one with current \(I_k \cdot (1 - \alpha)\), with \(I_k\) being the compressed current obtained from the envelope in analysis band \(k\), and \(\alpha\) being the current steering coefficient \((0 \leq \alpha \leq 1)\) derived from the spectral peak locator. Each analysis band \(k\) \((k = 1 \ldots N)\) is stimulated sequentially [see Sequential stimulation panel in Figure 3 (C)], completing a stimulation cycle. The Advanced Bionics CI has 16 electrodes and the F120 uses \(N = 15\) analysis bands.

Figure 3 (A) provides the concept of monopolar stimulation with its associated voltage spread. Figure 3 (B) demonstrates the concept of current steering (virtual channel) stimulation. With Paired and Triplet stimulation [Figure 3 (C)], each pulse is extended with zero stimulation after the end of the second biphasic pulse to keep the stimulation rate on each channel constant across sound coding strategies.
Figure 3. Monopolar (panel A) and current steering stimulation principle (panel B) with the resulting voltage spread. Panel C: Exemplary stimulation cycle for the F120 Sequential, Paired, and Triplet strategies.

2. Excitation Patterns using Sequential, Paired and Triplet Stimulation

The effect of spectral smearing using the F120 Sequential, Paired and Triplet strategies was first analyzed in simulation using a 3D finite-element model of the electrically-stimulated cochlea (see Nogueira, Schurzig, Büchner, Penninger, & Würfel, 2016 for details). Figure 4 demonstrates the output from this model as follows.

The spread of electric current in the cochlea is simulated in a 3D finite element method (FEM) from the geometry of the cochlea containing the Scala tympani, Scala vestibuli, Reisner membrane, basilar membrane, the modiolus and the nerve. The left panel in Figure 4 demonstrates the geometry of the auditory nerve. A spline interpolation of the auditory nerve compartment was used to create 10,000 nerve fibers along the cochlea. The 3D computer assisted drawing (CAD) model was generated in Inventor® (Autodesk, San Rafael, CA) and imported into COMSOL Multiphysics® (COMSOL Inc., Burlington, MA) to generate a tetrahedral mesh using the general physics algorithm. An electrode carrier with 16 half-band electrode contacts modelling the HiFocus 1J was created, as shown by the blue array in the left
panel of Figure 4. The physiology of the auditory nerve fiber was modeled as in Ashida and Nogueira (2018). The voltage distribution from the FEM, as shown in the right panel of Figure 4, was sampled at the most peripheral node of the nerve section. For each nerve fiber, the activation function in the most peripheral nerve was computed as in Eq. 4 from Nogueira et al. (2016). The currents delivered to each electrode across time, also known as electrodograms, for Sequential, Paired and Triplet stimulation were computed using the F120 sound coding strategy as described in Nogueira et al. (2009). Next the voltage distribution created by the electrodograms was estimated using the 3D voltage distribution model (see right panel of Figure 4). Finally, the neural excitation patterns, i.e. the neural activity across time, were computed using the nerve fiber model described in Nogueira et al. (2016) which is very similar to the one presented by Litvak et al. (2007).

Figure 4. *Left panel:* Geometry of the auditory nerve (red) with a 16-contact electrode array (blue); *Right panel:* 3D-voltage distribution (electric potential) model based on finite element method. Figure reproduced from (Nogueira et al., 2016) with permission.

German tokens were processed with the Sequential, Paired and Triplet sound coding strategies using the same levels of stimulation (threshold and most comfortable levels). The stimulation patterns served as input to the computational model that estimated the excitation patterns. Figure 5 presents the excitation patterns (number of spikes across fiber number and time) for
the German tokens /da/ and /go/ with the Sequential, Paired and Triplet strategies. Figure 5 demonstrated that increasing the number of parallel channels causes a clear spectral smearing of the excitation patterns.

![Neural excitation patterns for the token /da/ using the F120 Sequential (left column), Paired (middle column) and Triplet (right column) sound coding strategies. The rows indicate different voice cue manipulations. Top row: token spoken by original female speaker; Second row: neural excitation pattern after decreasing F0; Third row: neural excitation pattern after elongating VTL; Bottom row: neural excitation pattern after both decreasing F0 and elongating VTL.](image)

For each participant in the study, the F120 Sequential was fitted adjusting the most comfortable level of each electrode individually. Next, the strategy was activated, and the participant asked to loudness scale the presentation of a CCITT noise at 65 dB SPL free-field. The strategy was adjusted globally until the participant stated a comfortable loudness. Afterwards, the Paired or the Triplet sound coding strategies were fitted by globally adjusting the most comfortable level
across all electrodes by the same amount starting from the Sequential map fitting while presenting the same noise signal.

Figure 6. Current reduction in dB when fitting the Paired (left) or Triplet (right) strategies relative to Sequential to achieve the same loudness percept. The boxes extend from the lower to the upper quartile, and the middle line shows the median. The whiskers show the range of the data within 1.5 times the inter quartile range (IQR). Diamond-shaped symbols denote the mean.

Figure 6 shows the current reduction in dB for Paired and Triplet relative to Sequential stimulation across all participants recruited in this study. The plot demonstrates that the Sequential strategy requires higher currents than either the Paired or the Triplet strategies to elicit the same loudness percept, and that the Paired strategy requires higher current levels than the Triplet to reach most comfortable loudness, as was demonstrated by Langner et al., (2017). This is mainly due to the electrical interactions between the simultaneously stimulating channels, decreasing the necessary current for the same loudness. These interactions are
depending on the number of and the distance between the stimulating channels. The channel stimulation rate was kept constant across strategies by introducing a non-stimulating zero-phase after the end of the second phase of the biphasic pulse (see Figure 3b). This also implies the possibility to reduce power consumption even more, since an increase in the pulse duration requires much lower current levels to achieve the same loudness percept (Shannon, 1985, 1989) due to the resulting additional spread of excitation (McKay & McDermott, 1999). From this analysis, it can be concluded that adding parallel channels causes current smearing which, in turn, causes a reduction in the current levels required to achieve the same loudness percept, thus achieving the proposed current savings.

D. Task 1: F0 and VTL JNDs

1. Stimuli

The methods for this experiment are largely similar to the ones described in El Boghdady et al. (2019) and identical to those in El Boghdady et al. (in press). Therefore, they are described briefly here. Speech material from the Freiburg monosyllabic word test (Hahlbrock, 1953), which consisted of meaningful German monosyllabic words, were re-recorded for this study from an adult native German female speaker. The voice of the speaker had an estimated average F0 of 233 Hz and VTL of 13.9 cm based on her height (164 cm) using the data from Fitch and Giedd (1999). All recordings were equalized in root mean square (RMS) intensity.

Recordings were made in a sound-isolated anechoic chamber at the University Medical Center Groningen, NL, using a RØDE NT1-A microphone mounted on a RØDE SM6 with a pop-shield (RØDE Microphones LLC, CA, USA). The microphone was connected to a PreSonus TubePre v2 amplifier (PreSonus Audio Electronics, Inc., LA, USA) with noise filtering below 80 Hz. The amplifier output was recorded through the left channel of a DR-100 MKII TASCAM recorder (TEAC Europe GmbH, Wiesbaden, Germany) at a sampling rate of 44.1 kHz. Seventy-five consonant-vowel (CV) syllables were manually extracted from the recorded words in the
corpus, resulting in a list of combinations of the consonants [b, d, f, g, h, k, l, m, n, p, r, z, t, v, x, ts] and vowels [iː, oː, uː, a, ɛ, ɪ, ʊ, ɔ, eː].

A single trial consisted of concatenating three random CV syllables, with a 50-ms silence in between, to form a triplet of syllables. Within the trial, the same triplet of syllables was presented three times, with a 250 ms silence gap between each presentation. One of these three presentations was processed to have a different voice (lower F0, longer VTL, or both), as indicated by the vectors from the origin of the [$\Delta$F0, $\Delta$VTL] plane to the red crosses shown in Figure 1. All three presentations were resynthesized with STRAIGHT (Kawahara & Irino, 2005), even when F0 and VTL were not manipulated. The task was to select the triplet that had a different voice with respect to the other two in an adaptive 3-interval, 3-alternative forced choice task (3I-3AFC).

2. Procedure

Following the paradigm used in a number of previous studies (El Boghdady, Başkent, & Gaudrain, 2018; El Boghdady et al., 2019, e.g., in press; Gaudrain & Başkent, 2015, 2018), JNDs in this experiment were measured along three voice vectors, as indicated by the red crosses in Figure 1, using a 2-down 1-up adaptive procedure. This adaptive procedure results in 70.7% correct responses on the psychometric function (Levitt, 1971). A JND measurement consisted of a number of trials: a trial started with the target (voice-manipulated) triplet having a difference of 12 st relative to the other two reference triplets. After the participant’s response, a new trial began with a triplet composed of different combinations of syllables than the previous trial. If the participant was able to detect the voice manipulated triplet correctly on two consecutive trials, the voice difference between the reference triplets and the voice-manipulated triplet was reduced by 4 st. Otherwise, if the participant was unable to correctly identify the voice-manipulated triplet, the difference between the reference triplets and the voice-manipulated triplets was increased by the same step size. If the difference between the voice-
manipulated and reference triplets became less than twice the step size, the step size was reduced by a factor of $\sqrt{2}$. The procedure terminated after 8 reversals and the JND was calculated as the mean of the last 6 reversals.

The JND measurement for each of the three voice vectors was repeated three times per strategy, resulting in a total of 27 experimental conditions (3 voice vectors $\times$ 3 repetitions each $\times$ 3 coding strategies). Experimental conditions were blocked per strategy, meaning that a participant would perform all conditions for a given strategy before switching to the next one, and the order of the strategies was randomized per participant. Participants were blinded to the strategies tested.

Training was administered before the beginning of each strategy block with two voice spokes different than those used for data collection: $[\Delta F_0 = +5 \text{ st}, \Delta VTL = -7 \text{ st}]$ and $[\Delta F_0 = -12 \text{ st}, \Delta VTL = +3.8 \text{ st}]$. Each training condition was terminated after 6 trials, whether the algorithm had converged or not. Visual feedback was always provided.

### E. Task 2: Speech-on-Speech Intelligibility

#### 1. Stimuli

Stimuli taken from the German HSM sentence test (Hochmair-Desoyer et al., 1997) were used for the SoS intelligibility task, which is composed of 30 lists with 20 sentences taken from everyday speech, including questions. Sentences in this corpus are made up of three to eight words, with a single list containing 106 words in total. Lists 1-19 were used in this experiment and were previously recorded at the MHH from an adult native German female speaker, who had an average F0 of 218 Hz. All recordings were equalized in RMS intensity.

Four different masking voices were created as shown in Figure 1: the same talker as the target female [resynthesized with $\Delta F_0 = 0 \text{ st}, \Delta VTL = 0 \text{ st}$], a talker with a lower F0 relative to the
target female [$\Delta F_0 = -12 \text{ st}, \Delta VTL = 0 \text{ st}$], a talker with a longer VTL relative to the target female [$\Delta F_0 = 0 \text{ st}, \Delta VTL = +3.8 \text{ st}$], and a talker with both a lower F0 and a longer VTL relative to the target female to obtain a male-like voice [$\Delta F_0 = -12 \text{ st}, \Delta VTL = -3.8 \text{ st}$]. These conditions are referred to as Same Talker, F0, VTL, and F0+VTL, respectively, in the rest of this manuscript. The parameters for F0 and VTL were chosen based on the findings of an earlier study, in which CI users showed reduced SoS intelligibility and comprehension when the voice of the masker was manipulated with parameters taken from the top-right quadrant in Figure 1 (El Boghdady et al., 2019). In that study, the authors reasoned that masking voices taken from the lower-left quadrant, as performed in the current study, should be expected to yield a benefit in SoS performance for CI users. This premise was statistically tested in the results section of this experiment.

Test sentences were taken from lists 1-8 and 16-19, while maskers were constructed from lists 9 and 10. Training sentences were obtained from lists 11, 12, and 13, with one list randomly assigned per strategy. All sentences assigned for constructing the maskers were processed offline before data collection using STRAIGHT, with all combinations of $\Delta F_0$ and $\Delta VTL$ highlighted above. For the Same Talker condition, the masker sentences were also processed with STRAIGHT, without changing F0 or VTL. All target sentences were not processed with STRAIGHT.

Within a trial, the masker sequence started 500 ms before the onset of the target sentence and ended 250 ms after the offset of the target. For the specific $\Delta F_0$ and $\Delta VTL$ combination within the trial, the masker was constructed from random 1-second-long segments selected from the masker sentences previously processed with STRAIGHT. A raised cosine ramp of 2 ms was applied to the beginning and end of each segment before concatenating them to form the masker sequence. Finally, both the beginning and end of the entire masker sequence were ramped using a 50-ms raised cosine ramp.
Target sentences were calibrated at 65 dB SPL, and the intensity of the masker sequence was adjusted relative to that of the target to obtain the required target-to-masker ratio (TMR). The TMRs used for training and data collection in this task were set to +8 dB and +12 dB, respectively, following the protocol of El Boghdady et al. (2019). In that study, the authors demonstrated that a TMR of +8 dB has the potential of capturing group performance in the middle of the psychometric function (away from floor and ceiling effects). The stimuli for all three experiments were sampled at 44.1 kHz, processed, and presented using MATLAB R2014b (The MathWorks, Natick, MA).

2. **Procedure**

The SoS paradigm for this experiment was based on that used by El Boghdady et al. (2019, in press). A given trial consisted of presenting a single target-masker combination and the participant was asked to repeat what they heard from the target sentence. As in task 1, experimental conditions were blocked per strategy and the order of the strategies was randomized per participant.

A short training was provided for each strategy block, with both auditory and visual feedback. During the training phase of a given strategy, 12 sentences were randomly selected from the assigned training list: 6 sentences were presented in quiet, while the remaining 6 were presented with a competing masker. The masker voice used for training was assigned different values for $\Delta F_0$ and $\Delta VTL$ than those used during data collection (-6 st and +6 st, respectively).

Data collection was composed of a total of 240 trials for all three strategy blocks (20 sentences per list × 4 voice conditions × 3 strategies) generated offline prior to the beginning of the experiment. The trials within a strategy block were pseudo-randomized. No feedback was provided, and the stimulus was presented once. The participants’ responses were scored on a word-by-word basis using a graphical user interface (GUI) programmed in MATLAB.
Additionally, the verbal responses were recorded and stored as data files for later offline inspection.

Response words were scored in the following fashion: the German HSM sentences include words that are hyphenated in the corpus, such as ‘Wochen-ende’ (weekend). These words, although written without the hyphen, are hyphenated in the HSM corpus to be scored separately. Only the part repeated by the participant was marked as correct. Additionally, the response word was also considered correct if a participant changed the order of the words in the sentence. A response word was considered incorrect if only a part of the word was repeated for words that are not hyphenated in the HSM corpus, such as saying ‘füllt’ when the word was ‘überfüllt’ (crowded). Additionally, confusion of adjective form, e.g. saying ‘keiner’ (‘not any’ as used with a masculine noun) instead of ‘keine’ (‘not any’ as used with a feminine noun), or confusing the Dativ with the Akkusativ article, e.g. confusing ‘der’ with ‘dem’ or ‘den’, was also considered incorrect. Confusion of verb tenses or incorrect verb conjugation was considered incorrect. A total of four scheduled breaks were programmed into the experiment script, however, participants were encouraged to ask for additional breaks whenever they felt necessary.

**F. Task 3: Speech-on-Speech Comprehension**

1. **Stimuli**

The voice conditions for the masker in this experiment were the same as defined in experiment 1. The masker sequence was created as described in experiment 1 from lists 9 and 10 from the HSM material. Target sentences were based on German translations of the Dutch sentence verification task (SVT) developed by Adank and Janse (2009) and designed to measure sentence comprehension accuracy and speed (RT). This corpus is composed of 100 pairs of sentences, with each pair composed of a true (e.g. Bevers bouwen dammen in de rivier [Beavers

...
build dams in the river]) and false version (e.g. *Bevers grooien in een moestuin* [Beavers grow in a vegetable patch]). All sentences are grammatically and syntactically correct.

Translation from Dutch to German and the evaluation was performed thoroughly by three independent native German speakers: two of those speakers were also fluent in Dutch, while the third had sufficient knowledge of the language (see El Boghdady et al., in press, for a full description of the translation procedure). One sentence pair lost its meaning when translated to German and was discarded from the translations, resulting in 99 true-false sentence pairs. The additional four sentence pairs introduced by El Boghdady et al. (2019) for training purposes were translated to German as well.

Recordings were made in the same manner and using the same setup as those described in task 1. Recordings were taken from an adult native German female speaker, with an average F0 of 180 Hz, and an estimated VTL of about 14.1 cm following the method provided by Ives et al. (2005) and the data from Fitch and Giedd (1999).

### 2. Procedure

Following the paradigm in previous studies for the SVT (Adank & Janse, 2009; El Boghdady et al., in press; Pals, Sarampalis, Beynon, Stainsby, & Başkent, 2016), participants were asked if the target sentence was true (labeled ‘WAHR’) or false (labeled ‘UNWAHR’) by pressing the corresponding button on a button-box as quickly and accurately as possible within a time window of 6 seconds. The window was larger than the one used in Pals et al. (2016), to accommodate the CI users and not prime them to guess on most trials. If the time window was exceeded, the response was recorded as a no-response, and the next stimulus was presented. RTs were measured relative to the offset of the resolving word in the stimulus as was done by El Boghdady et al., (in press).
As was done in tasks 1 and 2, trials were also blocked per strategy, with the starting strategy randomized across participants. A short training was provided at the beginning of each strategy block. Twelve fixed sentence pairs were assigned for training and were excluded from data collection. Out of these pairs (24 true-false sentences), four true and four false sentences were randomly picked and assigned to the training block of each strategy. No true-false pair was assigned to the same training block.

In each training block, 2 true and 2 false sentences were first presented without a competing masker, followed by the remaining 2 true and 2 false sentences, which were presented with a competing masker. This masker also had the same voice parameters as those of the training masker voice used in task 2 and at the same training TMR of +12 dB. Both audio and visual feedback was provided: participants were shown if the sentence was true or false, and the sentence was shown on the screen while the whole stimulus was replayed through the loudspeaker.

The remaining sentences not used for training were used for data collection. These sentences were distributed among the number of conditions tested (4 masker voice conditions × 3 strategies), and no true-false pair was assigned to the same condition. All stimuli were generated offline for all three strategy blocks and pseudo-randomized within each block. During data collection, no feedback was given.

II. RESULTS AND DISCUSSION

All data were analyzed using R (version 3.3.3, R Foundation for Statistical Computing, Vienna, Austria, R Core Team, 2017), and regression models were implemented using the lme4 package (version 1.1-15, Bates, Mächler, Bolker, & Walker, 2015). When multiple comparisons were carried out, as in the case of the post-hoc analyses, a false-discovery rate (FDR) correction (Benjamini & Hochberg, 1995) was then applied to these p-values obtained from the multiple comparisons.
A. Task 1: Effect of channel interaction on F0 and VTL JNDs

Figure 7. JND distributions for F0, VTL, and F0+VTL cues obtained under each stimulation strategy: Sequential (dark grey boxes), Paired (light grey boxes), and Triplet (white boxes). F0: JNDs obtained along the negative F0 axis (lowering F0). VTL: JNDs obtained along the positive VTL axis (elongating VTL). F0+VTL: JNDs obtained along the diagonal with the combination F0 = -12 st, VTL = +3.8 st, simulating a male voice. The boxplot statistics are as indicated in Figure 6. The horizontal dashed line indicates a VTL difference of 3.8 st as used in the masker setup of tasks 2 and 3. The horizontal dotted line indicates an F0 difference of 12 st as used in the masker setup of tasks 2 and 3.

Figure 7 shows the JND distributions across all participants obtained for each voice cue, and indicates, as expected, a trend of worsening (increasing) JNDs as the amount of channel interaction increases (going from Sequential stimulation to Paired to Triplet). To investigate the general effect of channel interaction (stimulation strategy) on voice cue JNDs, a linear mixed-
effects model (LMM) was fitted to the log-transformed JNDs. This transformation was performed because the raw JNDs are bounded by zero and thus do not follow a normal distribution. The model was defined with strategy and voice cue (F0, VTL, and F0+VTL), along with their interaction, as the fixed-effect predictors. Interaction effects were included in the model to test whether the effect of strategy changes for different voice cues. Differences in baseline performance between participants, in addition to variations in the effect of strategy from one participant to the other, were accounted for in the linear model as random effects. To quantify the general effect of strategy on JNDs, a one-way type III repeated-measures analysis of variance (ANOVA) was applied to the aforementioned linear model and revealed a significant general effect of strategy on JNDs \[F(2,11.21) = 4.70, p = 0.03\], but no significant differences in JNDs between the different voice cues \[F(2,13.35) = 1.98, p = 0.18\]. The interaction effect between strategy and voice cue was also found to be non-significant \[F(4,55) = 0.91, p = 0.47\].

A similar LMM (including only a random intercept per participant as the random effect) was applied to each type of JND separately (F0, VTL, or F0+VTL) to study how stimulation strategy (channel interaction) affects each individual voice cue. A similar ANOVA to the one applied on the general model above was also applied here for each model separately, and \(p\)-values were then adjusted for multiple comparisons using the False Discovery Rate (FDR) method (Benjamini & Hochberg, 1995). These ANOVAs revealed that the general effect of strategy observed in the general model arose from a significant effect of strategy on F0 JNDs \[F(2,22) = 4.59, p = 0.03\] and F0+VTL JNDs \[F(2,22) = 4.56, p = 0.03\], but not on VTL JNDs \[F(2,22) = 1.23, p = 0.31\].

The post-hoc analyses of these tests revealed that F0 JNDs increased by about 1.44 st as the strategy changed from Sequential to Triplet \([\beta = 0.36, SE = 0.13, t(22) = 2.87, p = 0.03]\), but did not seem to be affected by Paired stimulation \([\beta = 0.07, SE = 0.13, t(22) = 0.59, p = 0.56]\).
On the contrary, VTL JNDS were neither affected by Paired [$\beta = 0.12$, $SE = 0.12$, $t(22) = 1.01$, $p = 0.39$] nor by Triplet stimulation [$\beta = 0.19$, $SE = 0.12$, $t(22) = 1.54$, $p = 0.21$] compared to Sequential. Finally, the participants’ JNDS to differences along both F0 and VTL (F0+VTL condition) also significantly increased (worsened) by about 1.35 st when the stimulation strategy was changed from Sequential to Triplet [$\beta = 0.37$, $SE = 0.12$, $t(22) = 3.02$, $p = 0.03$] but not from Sequential to Paired [$\beta = 0.19$, $SE = 0.12$, $t(22) = 1.57$, $p = 0.21$].

Taken together, these results indicate that when mild channel interaction exists, as was the case when Paired stimulation was compared to Sequential, sensitivity to voice cue differences was not significantly affected. However, as the channel interaction dramatically increased, as was the case with Triplet stimulation, sensitivity to voice cue differences was reduced. Because no significant interaction effect between stimulation strategy and voice cue was observed in the overall model, the effect of strategy should not be expected to differ for each voice cue. The fact that post-hoc analyses revealed no significant effect of strategy on VTL JNDS may have arisen from the relatively smaller differences in VTL JNDS across all three strategies compared to the F0 differences, even though a trend for worsening VTL JNDS could be observed. In an earlier study with vocoders, Gaudrain and Başkent (2015) have shown that when the number of effective spectral channels was sufficient, increasing channel interaction (shallower vocoder filters) did not lead to a significant worsening of VTL JNDS. Thus, an alternative explanation for these findings could be that the participants tested in the current study already had sufficient effective spectral channels which might have mitigated the detrimental effects of increased channel interaction.

A second observation concerns the effect of channel interaction on F0 JNDS. Because F0 information is encoded in both spectral and temporal cues (Carlyon & Shackleton, 1994), it was expected that the representation of F0 should have been robust to spectral degradations introduced by increased channel interaction. However, F0 cues were shown to be impaired by
increased channel interaction, indicating that the temporal aspect of these cues could not provide adequate F0 information for the CI listeners to reach the same JNDs as in the condition of minimal channel interaction (Sequential stimulation). These findings indicate that an adequate spectral resolution in the implant would be crucial for transmitting both F0 and VTL-related cues.

**B. Task 2: Effect of channel interaction on SoS intelligibility**

Figure 8. *Left panel:* SoS intelligibility score distribution across participants under each stimulation strategy (Sequential: dark grey; Paired: light grey; Triplet: white) for each masker voice condition. *Right panel:* same as left panel but demonstrating the effect of changing the masker voice for each stimulation strategy. *Same talker:* the condition when the target and masker were the same female speaker ($\Delta F0 = 0$ st, $\Delta VTL = 0$ st). *F0:* the condition when the masker had a lower F0 ($\Delta F0 = -12$ st, $\Delta VTL = 0$ st) relative to that of the target speaker. *VTL:* the condition when the masker had a longer VTL ($\Delta F0 = 0$ st, $\Delta VTL = +3.8$ st) relative to that of the target. *F0+VTL:* the condition when the masker had both a lower F0 and a longer VTL ($\Delta F0 = -12$ st, $\Delta VTL = +3.8$ st) relative to those of the target. The boxplot statistics are the same as described in Figure 6.
Figure 8 shows the distribution of SoS intelligibility scores across participants for each masker voice condition under each stimulation strategy. The scores in this figure were computed as the percentage of correctly repeated words out of the total number of words presented per condition. The data demonstrates that even though there is a large variability in performance across the CI participants for each stimulation strategy (left panel), there appears to be a trend for decreasing SoS intelligibility scores as the amount of channel interaction increases (going from Sequential stimulation to Paired to Triplet). In addition, the representation of the data in the right panel reveals that the degree of benefit in SoS intelligibility scores obtained from changing the masker voice relative to that of the target seems to decrease as the amount of channel interaction increases.

The binary per-word scores (0: incorrect; 1: correct) were modelled using logistic regression as implemented by a generalized linear mixed-effects model (GLMM) with a logit link function. The logistic regression model was fit to the binary per-word score with strategy and masker voice, along with their interaction, as the fixed-effects. The interaction between stimulation strategy and masker voice was included to test for the significance of the effect observed in the right panel of Figure 8, in which the degree of benefit in SoS intelligibility scores obtained from changing the masker voice seems to diminish as the amount of channel interaction increases. The GLMM was also defined to estimate a random intercept per participant to account for differences in baseline performance across participants. Additionally, random effects for strategy per participant and masker voice per participant were also included in the model to account for variations in the effect of strategy and masker voice on SoS intelligibility across participants.

As with the analyses of the JND task, an ANOVA (car package; Fox & Weisberg, 2011) was applied to the GLMM to test for the global effect of strategy, masker voice, and their interaction on the SoS intelligibility scores. Because this ANOVA is applied to a logistic regression model,
the output is a table of chi-squared ($\chi^2$) tests performed on the fixed-effects of the model instead of the traditional $F$-test statistics. The ANOVA revealed a significant effect of stimulation strategy [$\chi^2(2) = 27.29, p < 0.0001$], masker voice [$\chi^2(3) = 36.32, p < 0.0001$], and their interaction [$\chi^2(6) = 37.34, p < 0.0001$].

A post-hoc analysis was conducted using an ANOVA applied to the logistic regression model for the effect of strategy under each voice cue separately with FDR correction applied to the $p$-values. This analysis revealed that SoS intelligibility decreased as a function of increasing channel interaction for the Same Talker condition [$\chi^2(2) = 9.34, p = 0.01$], F0 condition [$\chi^2(2) = 8.99, p = 0.01$], VTL condition [$\chi^2(2) = 26.39, p < 0.0001$], and F0+VTL condition [$\chi^2(2) = 34.69, p < 0.0001$] (see left panel of Figure 8). These effects seemed to arise from the significant reduction in SoS intelligibility under Triplet stimulation compared to Sequential for most voice conditions [Same Talker: $\beta = -0.68$, SE = 0.23, $z = -2.92, p = 0.009$; F0: $\beta = -0.53$, SE = 0.28, $z = -1.91, p = 0.11$; VTL: $\beta = -1.00$, SE = 0.26, $z = -3.81, p < 0.001$; F0+VTL: $\beta = -1.03$, SE = 0.20, $z = -5.16, p < 0.0001$], but not between Paired and Sequential stimulation as obtained from the coefficients of the logistic regression model [Same Talker: $\beta = 0.05$, SE = 0.22, $z = 0.23, p = 0.82$; F0: $\beta = -0.10$, SE = 0.22, $z = -0.44, p = 0.75$; VTL: $\beta = -0.12$, SE = 0.26, $z = -0.45, p = 0.75$; F0+VTL: $\beta = -0.38$, SE = 0.24, $z = -1.59, p = 0.18$]. Consistent with the observations made from the JND task, a reduction in SoS intelligibility was observed with increasing channel interaction for all voice conditions. Thus, as channel interaction increases, spectral features that are important for both voice cue perception and SoS intelligibility appear to be degraded.

The significant interaction effect from the global ANOVA indicates that the benefit in SoS intelligibility obtained from changing the masker voice cues relative to those of the target was affected by the amount of channel interaction: as the channel interaction increased (going from Sequential stimulation to Paired to Triplet), the benefit obtained from the voice differences between masker and target speakers (going from Same Talker to F0 to VTL and then to
F0+VTL) decreased significantly (see right panel of Figure 8). In the Triplet case, the SoS intelligibility score for the baseline condition (Same Talker) was severely reduced compared to the same condition under Sequential stimulation. In addition, the largest benefit obtained from the condition F0+VTL under Triplet stimulation is almost the same as the mean intelligibility score for the same talker condition under either Sequential or Paired stimulation.

These findings reveal that substantial channel interaction may sufficiently degrade the signal to the extent that a benefit in SoS intelligibility from voice cue differences between two concurrent speakers may be impaired. Moreover, consistent with what has been observed in the JND task, CI participants appear to withstand mild channel interaction without a significant drop in their performance levels. However, as the channel interaction becomes more substantial, as is the case when Triplet stimulation is applied, overall SoS intelligibility scores start decreasing dramatically.

**C. Task 3: Effect of channel interaction on SoS comprehension accuracy and response times**

![Figure 9](image-url)

Figure 9. SoS comprehension accuracy in $d'$ (left panel), RT (middle panel), and drift rate (right panel) for each masker voice condition under each stimulation strategy. Boxplot statistics and description of conditions are the same as those described in the caption of Figure 6.
Figure 9 shows the SoS comprehension performance for each masker voice under each stimulation strategy. The right panel shows the effect of strategy on SoS comprehension accuracy converted to the sensitivity measure $d'$, computed as the ratio between the hit rate and the false alarm rate (Green & Swets, 1966). The $d'$ measure was used instead of percent correct because the $d'$ is unbiased to a participant’s particular preference for one response at the expense of the other. The large inter-participant variability appears to dilute the effect of strategy. As with the analyses applied to the data of the previous two tasks, an LMM was fit to the $d'$ data with strategy, masker voice, and their interaction as the fixed effects, and a random intercept per participant. Adding random slopes for the effect of strategy per participant and masker voice per participant did not improve the model fit to the data [$\chi^2(20) = 15.58, p = 0.74$], and was thus not included in the final LMM. An ANOVA similar to that applied to the LMM in the JND task was also applied to the LMM modeling the $d'$ data and revealed no effect of strategy [$F(2,77) = 2.68, p = 0.07$], masker voice [$F(3,77) = 1.82, p = 0.15$], or their interaction [$F(6,77) = 1.20, p = 0.31$] on the $d'$ accuracy scores.

The middle panel of Figure 3 shows the RT distributions obtained for each masker voice condition under each of the three stimulation strategies. Again, because of the large inter-participant variability, the effect of strategy on RTs is not evident. Because the RTs considered were those corresponding to only the correct responses, the number of RT data points differed across participants and conditions, which rendered the use of an ANOVA inappropriate. Additionally, the RT distributions per participant per condition were largely positively skewed. For these reasons, a GLMM with an inverse Gaussian distribution and inverse link function was fit to the RT data, as was suggested by Lo and Andrews (2015), and as was carried out by El Boghdady et al. (2019, in press). The GLMM best fitting the RT data included strategy, masker voice, and their interaction as the fixed-effects, in addition to random intercepts per participant. Including a random slope for strategy and masker voice per participant did not improve the overall model fit [Akaike information criterion (AIC) = 4213.03 and Bayesian information.
criterion (BIC) = 4362.18 for the model with random slopes versus AIC = 4205.78 and BIC = 4267.19 for the model without random slopes]. An ANOVA applied to the GLMM best fitting the RT data did not reveal either any effect of strategy [$\chi^2(2) = 0.006, p = 0.997$], masker voice [$\chi^2(3) = 0.049, p = 0.997$], or their interaction [$\chi^2(6) = 0.167, p = 0.9999$] on RTs. In this task, the no effect of strategy could be observed either for SoS comprehension accuracy or RTs when each measure was considered in isolation. Qualitatively, this implies that participants may be compromising accuracy for speed or vice versa, and that these response strategies differ per condition. Consider, for example, the $d'$ accuracy scores and RT data for the VTL condition. It appears that as participants give less accurate scores as the channel interaction increases, they also give these responses faster. However, this response strategy seems to change for the condition F0+VTL. In that condition, as channel interaction increases, participants also give less accurate responses, but they do so at slower speeds.

Thus, to address this speed-accuracy trade-off, a more unified measure of performance called the drift rate is helpful (for a review, see Ratcliff, Smith, Brown, & McKoon, 2016), and would be more suitable for assessing changes in difficulty level across the different conditions (Wagenmakers, Van Der Maas, & Grasman, 2007). The drift rate, as shown in the right panel of Figure 9, estimates the participants' accuracy rate for each condition, and was computed using the EZ-drift diffusion model (EZ-DDM; Wagenmakers et al., 2007), which is a simplified version of the full model proposed by Ratcliff (1978). It was not possible to fit full diffusion models, such as that provided in the R package diffIRT (Molenaar, Tuerlinckx, van der Maas, & others, 2015) to the data from this task because the nature of the paradigm used introduced entries with missing data, as described in the methods section. In this situation, the simplified EZ-DDM model was more appropriate. The EZ-DDM makes three assumptions about the data: 1) the RT distributions are positively skewed for each condition (strategy and masker voice); 2) There are no differences between the RT distributions of correct and incorrect responses for each participant under each condition;
3) No significant interaction should be present between response accuracy and stimulus category (true/false categories). Following the method proposed by Wagenmakers et al. (2007) for checking whether these assumptions are met by the data, the check performed on the first assumption revealed that the RT distributions were positively skewed.

For assumptions 2 and 3, because they were required to be checked for each participant under every condition, there were a few participants who failed the checks for a number of conditions, as was also the case with the data analyzed by Wagenmakers et al. (2007). For assumption 2, P06 failed this check for 2 conditions, P07 and P08 failed this check for one condition, and P10 and P11 failed this check for 4 conditions. This indicates that for these conditions, the participants may have had systematically slower or faster error responses compared to correct responses.

For assumption 3, only P07, P08, P10, and P11 failed this check for one condition. This indicates that for this condition, P07, P08, P10, and P11 may have had some bias to classify a stimulus as true or false. Nonetheless, the EZ-DDM was still applied here to obtain an approximation for the drift rate. This was motivated by the fact that the sample data used by Wagenmakers et al. also violated the aforementioned assumptions to a larger degree than the data presented in the current manuscript, however, Wagenmakers et al. demonstrated that the EZ-DDM was still able to provide reasonable estimates to the drift rate that were in line with those obtained from the full diffusion model.

An LMM was fit to the drift rate data, with strategy, masker voice, and their interaction as fixed effects, and with random intercepts estimated per participant as the random effect. Attempting to include a random slope for strategy per participant and masker voice per participant yielded a non-converging model, and thus were not included in the final LMM. An ANOVA applied to the final LMM revealed a significant effect of strategy [$F(2,70.21) = 4.30, p = 0.02$] and masker voice [$F(3,70.15) = 3.11, p = 0.03$] on the drift rate, but not of the interaction term on the drift rate [$F(3,70.10) = 0.84, p = 0.54$]. This indicates that as the channel interaction increases, and
as the voice difference between masker and target increases, the accuracy rate on the SVT task appears to decrease, which is in line with the results observed in task 2.

III. CONCLUSION

This study investigated whether increasing channel interaction as a result of simultaneously stimulating multiple channels in the CI would lead to a reduced sensitivity to F0 and VTL cues (task 1), and correspondingly, reduced SoS intelligibility and comprehension performance (tasks 2 and 3). The data from task 1 — JND — revealed that, in line with what was expected, increasing channel interaction significantly reduced CI users’ sensitivity to voice cues (both spectral and temporal features), as demonstrated by the main effect of stimulation strategy in addition to a lack of interaction effect between voice cue and stimulation strategy.

The data from task 2 — SoS intelligibility — demonstrated an effect of channel interaction, a benefit from voice differences between target and masker speakers, and a significant interaction effect between these two factors. Compared to Sequential stimulation, increasing the channel interaction was shown to impair SoS intelligibility scores only for Triplet stimulation but not for Paired stimulation. This indicates that for mild cases of channel interaction baseline SoS intelligibility could still be maintained. However, for more extreme cases of channel interaction, as in the case of Triplet stimulation, SoS intelligibility scores become severely degraded. In addition, the lack of detrimental effect of Paired stimulation on voice cue sensitivity and SoS intelligibility provides evidence that parallel stimulation could be utilized as a method for reducing power in CIs without impairing performance on tasks relying on voice cue perception.

The voice parameters for F0 and VTL assigned for the maskers in this study (starting from a female voice and approaching a male-like voice) yielded a benefit in SoS intelligibility. In a previous study (El Boghdady et al., 2019), the authors demonstrated that voice parameters taken from the top-right quadrant of the [ΔF0, ΔVTL] plane (towards child-like voices) failed to provide release from masking for CI users, even though the differences between those child-
like voices and the reference female speaker were larger than those between the male-like voices and the reference female speaker in the current study. Taken together, these data indicate that CI users may benefit differently from voice cue differences depending on which speaker space they cover. However, this benefit from voice differences between target and masker is reduced as the amount of channel interaction increases, as was demonstrated by the significance of the interaction effect observed between stimulation strategy and voice cue. This means that as channel interaction becomes substantial, CI listeners may not be able to benefit from voice cue differences between competing talkers in SoS scenarios.

The data from task 3 — SoS comprehension — also corroborated the findings from the previous two tasks regarding the effect of channel interaction. Although comprehension accuracy and RT measures revealed no effect of either channel interaction or voice cue, the drift rate measure was able to demonstrate a detrimental, albeit small, effect of increased channel interaction and masker voice. These findings support the idea that drift rate as a combined measure of comprehension accuracy and speed may provide insight into the data that may not be initially visible in either accuracy or RT measures alone.

The findings from this study highlight the importance of spectrotemporal resolution when performing tasks that depend on voice-cue perception. This raises the question of whether CIs could be fitted with the goal of mitigating the effect of decreased spectrotemporal resolution that may arise from channel interaction. Several studies (e.g., Di Nardo, Scorpecci, Giannantonio, Cianfrone, & Paludetti, 2011; El Boghdady et al., 2018; Fitzgerald, Sagi, Morbiwala, Tan, & Svirsky, 2013; Fu & Shannon, 1999; Grasmeder, Verschuur, & Batty, 2014; Leigh, Henshall, & McKay, 2004; McKay & Henshall, 2002, 2002; Omran, Lai, & Dillier, 2011) have proposed that optimizing the frequency-to-electrode allocation map could have the potential to address the limited spectral resolution in the implant. More specifically, using vocoder simulations, El Boghdady et al. (2018) have shown that the frequency-to-electrode
allocation map could have a direct influence on VTL JNDs, and that the frequency mapping, if optimally fitted, could help reduce the detrimental effects of channel interaction and frequency mismatch in the cochlea on VTL JNDs. These studies help to pave the way for investigating whether the CI parameters (such as the frequency allocation map) or signal processing could be optimized in a way to improve both SoS perception and the sensitivity to voice cues.

ACKNOWLEDGEMENTS

The work presented here was jointly funded by Advanced Bionics (AB), the University Medical Center Groningen (UMCG), the PPP-subsidy of the Top Consortia for Knowledge and Innovation of the Ministry of Economic Affairs, and the DFG Cluster of Excellence EXC 1077/1 “Hearing4all”. The study was additionally supported by a Rosalind Franklin Fellowship from the University Medical Center Groningen, University of Groningen, and the VICI Grant No. 016.VICI.170.111 from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development (ZonMw). This work was conducted in the framework of the LabEx CeLyA (“Centre Lyonnais d’Acoustique”, ANR-10-LABX-0060/ANR-11-IDEX-0007) operated by the French National Research Agency, and is also part of the research program of the Otorhinolaryngology Department of the University Medical Center Groningen: Healthy Aging and Communication. Waldo Nogueira and Florian Langner were funded by the DFG Cluster of Excellence EXC 1077/1 “Hearing4all”. The authors would like to especially thank: Eugen Kludt and the rest of the MHH research group for their support; Luise Wagner, Annika Luckman, Anita Wagner, Alana Wulf, Enja Jung, Olivier Crouzet, Charlotte de Blecourt, Fergio Sismono, and Britt Bosma for their help setting up the German SVT material, in addition to the speakers who recorded the German SVT material; the CI participants who took part in this study.
REFERENCES


