Effect of Spectral Contrast Enhancement on Speech-on-Speech Intelligibility and Voice Cue Sensitivity in Cochlear Implant Users

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

Effect of Spectral Contrast Enhancement on Speech-on-Speech Intelligibility and Voice Cue Sensitivity in Cochlear Implant Users

Nawal El Boghdady,[1,2] Florian Langner,[3] Etienne Gaudrain,[4,1,2] Deniz Başkent,[1,2] and Waldo Nogueira[3]

Objectives: Speech intelligibility in the presence of a competing talker (speech-on-speech; SoS) presents more difficulties for cochlear implant (CI) users compared with normal-hearing listeners. A recent study implied that these difficulties may be related to CI users' low sensitivity to two fundamental voice cues, namely, the fundamental frequency (F0) and the vocal tract length (VTL) of the speaker. Because of the limited spectral resolution in the implant, important spectral cues carrying F0 and VTL information are expected to be distorted. This study aims to address two questions: (1) whether spectral contrast enhancement (SCE), previously shown to enhance CI users' speech intelligibility in the presence of steady state background noise, could also improve CI users' SoS intelligibility, and (2) whether such improvements in SoS from SCE processing are due to enhancements in CI users' sensitivity to F0 and VTL differences between the competing talkers.

Design: The effect of SCE on SoS intelligibility and comprehension was measured in two separate tasks in a sample of 14 CI users with Cochlear devices. In the first task, the CI users were asked to repeat the sentence spoken by the target speaker in the presence of a single competing talker. The competing talker was the same target speaker whose F0 and VTL were parametrically manipulated to obtain the different experimental conditions. SoS intelligibility, in terms of the percentage of correctly repeated words from the target sentence, was assessed using the standard advanced combination encoder (ACE) strategy and SCE for each voice condition. In the second task, SoS comprehension accuracy and response times were measured using the same experimental setup as in the first task, but with a different corpus. In the final task, CI users' sensitivity to F0 and VTL differences were measured for the ACE and SCE strategies. The benefit in F0 and VTL discrimination from SCE processing was evaluated with respect to the improvement in SoS perception from SCE.

Results: While SCE demonstrated the potential of improving SoS intelligibility in CI users, this effect appeared to stem from SCE improving the overall signal to noise ratio in SoS rather than improving the sensitivity to the underlying F0 and VTL differences. A second key finding of this study was that, contrary to what has been observed in a previous study for childlike voice manipulations, F0 and VTL manipulations of a reference female speaker (target speaker) toward male-like voices provided a small but significant release from masking for the CI users tested.

Conclusions: The present findings, together with those previously reported in the literature, indicate that SCE could serve as a possible background-noise-reduction strategy in commercial CI speech processors that could enhance speech intelligibility especially in the presence of background talkers that have longer VTLs compared with the target speaker.

Key words: Cochlear implant, Speech-on-speech, Spectral contrast enhancement, Voice, Pitch, Vocal tract length.

(Ear & Hearing 2021;42;271–289)

This article has received an OSF badge for Open Data.

INTRODUCTION

Understanding speech in the presence of background interference is quite challenging for cochlear implant (CI) users compared with normal-hearing (NH) listeners (e.g., Fu et al. 1998; Friesen et al. 2001; Stickney et al. 2004; El Boghdady et al. 2019). In such scenarios, a listener attempts to extract relevant spectrotemporal information from the target speech while trying to suppress interference from the background masker (for a review, see Assmann & Summerfield 2004; Brungart et al. 2006). NH listeners have been shown to utilize spectral dips or temporal modulations in fluctuating maskers to obtain higher target-speech intelligibility and release from masking (Duquesnoy 1983; Festen & Plomp 1990; Gustafsson & Arlinger 1994; Nelson et al. 2003; Cullington & Zeng 2008). Unmodulated (steady state) noise which is spectrally matched to the long-term average spectrum of the target speech (speech-shaped noise; SSN) was found to yield a larger masking effect in NH listeners compared with amplitude modulated (fluctuating) SSN (Nelson et al. 2003) and speech maskers (Turner et al. 2004; Cullington & Zeng 2008). On the contrary, CI users appear to make no use of such dips; modulations introduced in SSN maskers produced no release from masking (Nelson et al. 2003), while the competing speech was observed to be generally much worse than SSN (Stickney et al. 2004; Cullington & Zeng 2008).

The question thus arises, why would CI users, on average, find speech maskers to be more challenging than SSN while NH listeners mostly experience release from masking from speech

[1]Department of Otorhinolaryngology, University Medical Center Groningen, Groningen, the Netherlands; [2]Graduate School of Medical Sciences, Research School of Behavioral and Cognitive Neurosciences, University of Groningen, Groningen, the Netherlands; [3]Department of Otolaryngology, Medical University Hannover and Cluster of Excellence Hearing4all, Hanover, Germany; and [4]CNRS UMR 5292, INSERM U1028, Lyon Neuroscience Research Center, Universite de Lyon, Lyon, France.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and text of this article on the journal's Web site (www.ear-hearing.com).

<zdoi; 10.1097/AUD.0000000000000936>

maskers compared with SSN maskers? A possible explanation for these observations could be that NH listeners utilize voice differences between talkers in multi-talker settings (e.g., Brungart 2001; Stickney et al. 2004; Cullington & Zeng 2008; Darwin et al. 2003; El Boghdady et al. 2019); however, CI users seem to derive little or no benefit from such voice differences (Stickney et al. 2004; Cullington & Zeng 2008; El Boghdady et al. 2019). In fact, a recent study has shown that such speech-on-speech (SoS) perception in CI listeners is linked to their sensitivity to two principal voice cues, namely the fundamental frequency (F0) and the vocal tract length (VTL) of the speaker (El Boghdady et al. 2019). This study demonstrated that the lower the CI users' sensitivity to both of these two cues, the lower their overall SoS performance was. Yet unlike NH listeners, CI users, on average, did not benefit from the voice manipulation that increased F0 and VTL differences between the two competing talkers.

The speaker's F0 is responsible for the perception of the voice pitch and is usually higher for adult women than adult men (Peterson & Barney 1952; Smith & Patterson 2005). Such F0 cues can be encoded in not only the temporal envelope and the temporal fine structure (e.g., Moore 2008; Wang et al. 2011; Cabrera et al. 2014), but also the cochlear location of excitation (e.g., Licklider 1954; Carlyon & Shackleton 1994; Oxenham 2008). The speaker's VTL correlates with their physical size (Fitch & Giedd 1999) and hence is usually shorter for adult women than for adult men. The VTL provides the listener with cues regarding the speaker's size (Ives et al. 2005; Smith et al. 2005). Such cues are usually represented in the relationship between the locations of the spectral peaks (formants) and spectral valleys (Chiba & Kajiyama 1941; Müller 1848; Stevens & House 1955; Fant 1960; Lieberman & Blumstein 1988). Shortening VTL results in the stretching of the spectral envelope toward higher frequencies on a linear frequency scale while elongating VTL results in the compression of the spectral envelope toward lower frequencies. Thus, while F0 cues have a spectrotemporal representation, VTL cues are mainly spectral in nature. Hence, the adequate representation of these two cues would be expected to require sufficient spectrotemporal resolution, such that the relationship between the locations of the spectral peaks and valleys are adequately maintained. However, because of the limited spectrotemporal resolution of the implant (Fu et al. 1998; Nelson & Jin 2004; Fu & Nogaki 2005), the transmission of F0 and VTL cues in CI devices is expected to be impaired. This idea has been directly tested in vocoder simulations of CI processing with NH listeners so as to better control the parameters of the simulated spectrotemporal degradation. In one such study, Gaudrain and Başkent (2015) modeled the effective number of spectral channels and channel interaction as the number of vocoder channels and filter-slope shallowness, respectively. The authors showed that, in line with what is expected, as the number of spectral channels decreases, and as the channel interaction increases, the sensitivity to VTL cues deteriorates. Supporting these observations from vocoder studies, a later study by the same authors showed that, compared with NH listeners, CI users have particularly poor sensitivities to both F0 and VTL differences (Gaudrain & Başkent 2018; Zaltz et al. 2018), which is also in line with CI users' reported abnormal use of these voice cues, especially VTL, for gender categorization (e.g., Fuller et al. 2014; Meister et al. 2016) and SoS perception (Pyschny et al. 2011; El Boghdady et al. 2019). Thus, the poor spectrotemporal resolution in CIs is also expected to influence the utilization of voice differences between target and masker speakers in SoS scenarios.

Spectral contrast enhancement (SCE) algorithms, which attempt to improve the contrast between spectral peaks and valleys in the signal, have been proposed as a method for mitigating the detrimental effects of the limited spectrotemporal resolution in the implant. This has been supported by the observation that CI users require higher spectral contrast than NH listeners to correctly identify synthetic vowels (Loizou & Poroy 2001): a task that relies mainly on spectral resolution. To that end, SCE algorithms have been shown to provide small but significant improvements in speech intelligibility in steady state SSN maskers (e.g., Baer et al. 1993; Bhattacharya & Zeng 2007; Bhattacharya et al. 2011; Nogueira et al. 2016; Chen et al. 2018).

In one such study, Baer et al.'s (1993) SCE algorithm was shown to provide significant improvements in speech intelligibility in steady state SSN for listeners with hearing loss for moderate degrees of spectral enhancement. Later, Turicchia and Sarpeshkar (2005) proposed a compressing-expanding (companding) strategy; a strategy that attempts to simulate the two-tone suppression phenomenon and compression effects occurring in a biological cochlea. The authors' proposed companding strategy was observed to provide SCE as an emergent property, and thus the authors argued for its potential to improve speech intelligibility in background noise. The parameters for this companding strategy were fit in later studies and provided significant improvements in speech-in-noise intelligibility when tested with vocoder simulations of CI processing (Oxenham et al. 2007), or with NH and CI listeners (Bhattacharya & Zeng 2007; Bhattacharya et al. 2011). Yet, these aforementioned algorithms were implemented as a front-end stage that preprocessed all stimuli off-line before they could be processed through the CI speech processor. Such front-end processing blocks make it difficult to control the exact amount of SCE applied to the stimulus because the CI processing pathway contains multiple nonlinear operations, such as automatic gain control.

To address these aforementioned issues, a real-time implementation based on the algorithm from the study by Loizou and Poroy (2001) was developed in the study by Nogueira et al. (2016) as an extra processing stage in the signal processing pathway of the standard advanced combination encoder (ACE) strategy typically used in Cochlear Limited (Sydney, Australia) devices. Such an implementation provides better control for the amount of SCE applied to the stimuli and provides easier testing in real-time with CI users and was hence used in the present study. Consistent with the findings reported by Baer et al. (1993) for listeners with hearing loss, Nogueira et al. (2016) also showed that for moderate degrees of spectral enhancement, improvements in speech intelligibility in SSN were observed for CI users when the output from their SCE strategy was matched in loudness to that of the control ACE strategy. Yet, it remains unknown whether SCE can improve speech intelligibility when the target speaker is masked by another competing talker (SoS), a situation in which the perception of F0 and VTL cues could be crucial (Brungart 2001; Darwin et al. 2003; Başkent & Gaudrain 2016; El Boghdady et al. 2019).

On the basis of the findings of the aforementioned studies, the aim of this study was to investigate whether SCE could improve SoS perception and voice cue sensitivity in CI users. Two

research questions were posed: (1) whether SCE would enhance SoS perception in CI users and (2) whether this improvement would arise from SCE's enhancement of the underlying sensitivity to F0 and VTL differences between the target and masker speakers. The expectations were that these improvements from SCE should be larger for VTL compared to F0 perception, because VTL is a primarily spectral cue, while F0 is both spectral and temporal in nature. The first research question was addressed by experiments 1 and 2, which differed in the speech material and type of SoS test administered to the CI users. The aim of using more than one SoS test was two-fold. The first aim was to have two tasks that measure different aspects of speech perception, namely intelligibility and comprehension, which may also potentially differ in task difficulty, akin to the paradigm followed in an earlier study by El Boghdady et al. (2019). Experiment 1 measures word intelligibility in the context of meaningful sentences, while experiment 2 measures the overall sentence comprehension. The second aim was to avoid potential floor effects that might be observed in the intelligibility task. Because the target-to-masker-ratio was adjusted based on simulations and pilot runs with only a few participants, there was a risk that performance measured for the intelligibility task might still be around floor levels, especially with the large intersubject variability expected from CI users. In addition, participants usually anecdotally reported that they found the intelligibility task to be difficult. For these reasons, a second sentence comprehension task was included. The second hypothesis of this study, namely, whether SCE has the potential of improving CI users' sensitivity to F0 and VTL differences, was addressed in experiment 3.

MATERIALS AND METHODS

All experiments in this study were based on the paradigms from the study by El Boghdady et al. (2019). This section describes the methods common to all three experiments conducted in this study. Methods particular to a given experiment are described in detail under the heading of the corresponding experiment.

Participants

CI participants 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 higher than 20% in SSN at a 10-dB signal to noise ratio on the Hochmair-Shulz-Moser (HSM) sentence test (Hochmair-Desoyer et al. 1997). Because stimuli were presented in free-field, participants were also selected to have no residual acoustic hearing in the implanted ear (no thresholds better than 80 dB HL at all audiometric frequencies). In addition, all participants recruited had more than one year of CI experience and were all postlingually deafened. All participants were native German speakers and reported no health problems, such as dyslexia or attention deficit hyperactivity disorder.

Fitting these criteria, 14 CI users (6 male participants) 39 to 81 years old ($\mu = 63$ years, $\sigma = 13.3$ years) with Cochlear Nucleus devices volunteered to take part in this study. Not all 14 participants were able to complete all three experiments because of their difficulty: Participant P14 was only able to complete experiment 3 (voice just-noticeable-differences [JNDs]), while Participant P13 was only able to complete experiments 1 (SoS intelligibility) and 3 (voice JNDs). Thus, in total, out of the 14 CI participants, 13 took part in experiment 1, 12 took part in experiment 2, and all 14 took part in experiment 3. The participant demographics are reported in Table 1.

This study was approved by the institutional medical ethics committee of the MHH (protocol number: 3266-2016). All participants were given ample information and time to consider the study before participation and signed a written informed consent before data collection. Participation was entirely voluntary, but travel costs were reimbursed.

Voice Cue Manipulations

All voice manipulations were conducted relative to the original speaker of the corpus deployed in each experiment using the STRAIGHT (*Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram)* speech manipulation system (Kawahara & Irino 2005). F0 manipulations were realized by shifting the pitch contour of the entire speech stimulus by a designated number of semitones (12th of an octave; st). For increases in F0, the pitch contour is shifted upwards toward higher frequencies relative to the average F0 of the stimulus. For decreases in F0, the pitch contour is shifted downwards toward lower frequencies. VTL changes were implemented by expanding or compressing the spectral envelope of the signal: elongating/shortening VTL results in a compression/stretching of the spectral envelope toward lower/higher frequency components.

Figure 1 shows the [ΔF0, ΔVTL] plane, with the original female voice of the corpus used in experiment 1 placed at the origin of the plane (solid black circle). The dashed ellipses represent the ranges of relative F0 and VTL differences between the original female voice and 99% of the population according to the data from the study by Peterson and Barney (1952). The Peterson and Barney data were normalized here relative to the adult female speaker in the corpus used in experiment 1. This speaker had an average F0 of about 218 Hz and an estimated VTL of around 14 cm. The VTL was estimated following 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 reference female speaker based on published growth curves for the German population (Schaffrath Rosario et al. 2011; Bonthuis et al. 2012). Negative ΔVTLs denote a shortening in the VTL of the speaker and vice versa, thus ΔVTL is oriented upside down to indicate that negative ΔVTLs yield an increase in the frequency components of the spectral envelope of the signal. The red crosses indicate the four combinations of F0 and VTL differences used to create the masker speech in experiments 1 and 2, and 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). These particular values were chosen to address a potential question of whether CI users would demonstrate a benefit from voice differences along the male voice space since the data from the study by

TABLE 1. Demographic Information for CI Users

Participant Number	Age at Testing (yr)	Processor	Implant	Duration of Device Use (yr)	Duration of Hearing Loss (yr)	Etiology
P01	54	CP910	Nucleus_CI24RE (CA)	11.5	2.2	Unknown
P02	48	CP910	Nucleus_CI522	2.8	Progressive	Unknown
P03	73	CP910	Nucleus_CI512	8.4	0.8	Genetic
P04	81	CP910	Nucleus_CI24R (CS)	15.9	0.3	Sudden hearing loss
P05	77	CP910	Nucleus_CI24R (CS)	15.6	3.1	Sudden hearing loss
P06	66	CP910	Nucleus_CI422	7.9	Progressive	Unknown
P07	39	CP950 Kanso	Nucleus_CI512	8.1	Unknown	Congenital Rubella syndrome
P08	78	CP950 Kanso	Nucleus_CI24R (CS)	16.4	3.2	Sudden hearing loss
P09	48	CP810	Nucleus_CI24RE (CA)	6.2	Progressive	Unknown
P10	59	CP810	Nucleus_CI422	7.5	Progressive	Sudden hearing loss
P11	52	CP910	Nucleus_CI512	7.6	46.7	Otosclerosis cochleae
P12	64	CP910	Nucleus_CI24R (CA)	13.5	31.7	Unknown
P13	71	CP910	Nucleus_CI24RE (CA)	11.7	0.6	Unknown
P14	76	CP910	Nucleus_CI422	5.8	Progressive	Unknown

All durations in years are calculated based on the date of testing. Progressive hearing loss refers to minimal hearing loss that gradually progressed until the participant eventually fulfilled the criteria for acquiring a CI.
CI, cochlear implant.

El Boghdady et al. (2019) demonstrated no benefit from voice differences along the child voice space.

Signal Processing

Advanced Combination Encoder • The ACE strategy, shown in Figure 2, was selected as the reference strategy to which SCE was compared. The ACE strategy first digitizes the acoustic signal and applies automatic gain control, which compresses the large dynamic range of the input acoustic signal to the smaller dynamic range of the implant. The signal is then analyzed at a sampling frequency of 15,659 Hz using a sliding Hann window comprising a temporal frame of 128 samples. To each temporal frame, a fast Fourier transform is applied to decompose the acoustic audio signal into *M* frequency bands. Next, the envelope of each band is extracted, and the *N* bands with the highest amplitudes (*N* maxima) are selected from the available *M*, making ACE an *N-of-M* strategy. Finally, a loudness growth function is applied to the *N* selected bands to map their compressed amplitudes to the dynamic range specified by the participant's threshold (T) and comfort (C) levels, which are then converted to current units before stimulating the electrodes. Additional details on the ACE strategy can be found in the studies by Nogueira et al. (2005; 2016).

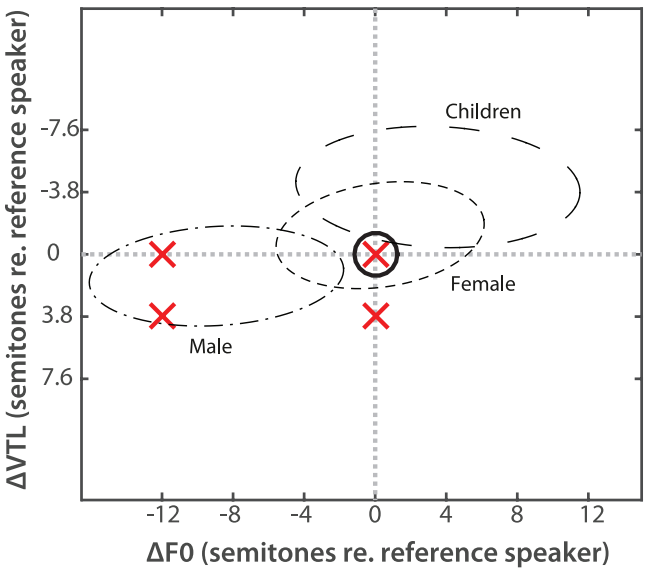

Fig. 1. [ΔF0, ΔVTL] plane, which represents the relative difference (Δ) in F0 and VTL between the reference female speaker from experiment 1 (indicated by the solid circle at the origin of the plane) and 99% of the population (dashed ellipses). Decreasing F0 and elongating VTL yields deeper-sounding male-like voices, while increasing F0 and shortening VTL yields childlike voices. The dashed ellipses are based on the Peterson and Barney data (1952), which were normalized to the reference female speaker. The red crosses indicate the four different combinations (experimental conditions) of ΔF0 and ΔVTL used in both experiments 1 and 2, and the voice vectors from the origin of the plane along which the JNDs were measured in experiment 3. JND indicates just-noticeable-difference; VTL, vocal tract length.

Spectral Contrast Enhancement • The SCE algorithm used in this study, as was implemented by Nogueira et al. (2016), first locates the three most prominent spectral peaks, where the formant frequencies are expected to lie, in addition to the valleys in between those peaks. The original spectral contrast between the selected peaks and their corresponding valleys in each frame is then determined as the difference between those peaks and valleys on a dB scale. Then, the desired enhancement (attenuation factor) to be applied to the entire spectral envelope, except for the three most prominent peaks, is computed by specifying a parameter in the algorithm called the SCE factor, such that for an SCE factor of 0, no enhancement is applied, which would result in the output of the reference ACE strategy. For an SCE factor of 1, the spectral contrast between the three most prominent peaks and their corresponding valleys is doubled on a dB scale, and for an SCE factor of 0.5, the spectral contrast is increased by a factor of 1.5 on a dB scale. This results in the preservation of the levels of the three most prominent peaks, the enhancement of the contrast between those peaks and their corresponding valleys, and the attenuation of the remaining peaks and valleys relative to the enhanced contrast between the three most prominent peaks and their valleys. The signal processing pathway then proceeds to select the *N* maxima. Figure 3 shows the effect of ACE and SCE processing, with multiple SCE factors, on a sample phoneme.

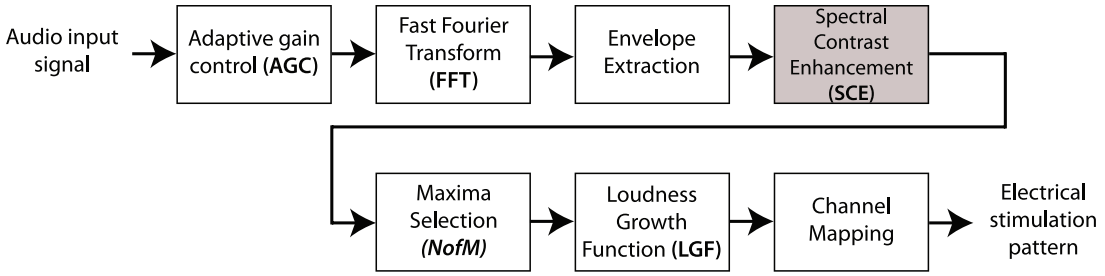

Fig. 2. Signal processing pathway for ACE and SCE. The pathway for SCE is identical to that of ACE, except for the addition of an extra processing block with the SCE algorithm (shaded block) before maxima selection. When the SCE factor is set to 0 (see paragraph on SCE processing), no SCE processing is applied, which results in the ACE processing strategy. Figure reproduced from Nogueira et al. (2016) with permission. ACE, advanced combination encoder; SCE, spectral contrast enhancement.

In a simulation, Nogueira et al. (2016) have shown that applying SCE before the maxima selection block influences the peak selection in a manner that reduces potential channel interaction when compared with ACE. Thus, it can be hypothesized that the reduced channel interaction should contribute to the enhanced overall spectral resolution which might improve the perception of spectrally-related voice cues, such as VTL.

Because SCE maintains the levels of the most prominent three peaks and attenuates the remaining peaks and valleys, sounds processed by SCE are softer than those processed by ACE (Nogueira et al. 2016). For this reason, to compare the two strategies, their perceived loudness should be equated, as was done by Nogueira et al. (2016). This loudness balancing procedure was also deployed in the present study and is described in detail in Procedure section. The stimuli for all three experiments were sampled at 44.1 kHz, processed, and presented using a custom-built script in MATLAB R2014b (The MathWorks, Massachusetts, USA).

Apparatus

Both the ACE strategy and appended SCE block were implemented in Simulink to run in real-time on a Speedgoat xPC target machine (Goorevich & Batty 2005). The experiment script implemented in MATLAB was run on a standard Windows computer and was responsible for stimulus delivery. All stimuli were calibrated to 65 dB SPL, which was the reference for the loudness balancing procedure as explained in detail in Procedure section. Stimuli were delivered through a Fireface 800 soundcard (RME, Haimhausen, Germany) connected to a Genelec 8240A loudspeaker (Genelec, Iisalmi, Finland) positioned 1 m from a Cochlear System 5 microphone array. This microphone is the same as that in the Cochlear clinical speech

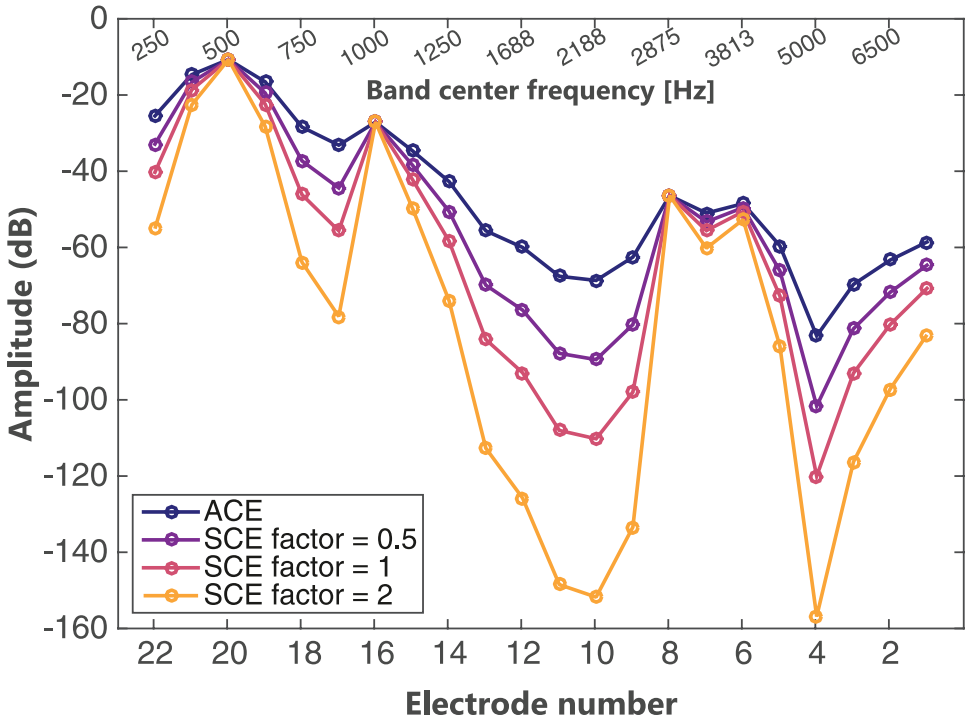

Fig. 3. Effect of ACE and SCE strategies on the spectral envelope of a single frame of the German vowel /o:/. The three most prominent spectral peaks are maintained while the valleys in between, and any subsequent peaks and valleys, are attenuated according to the SCE factor. Higher SCE factors denote larger spectral enhancements. Band numbers are in descending order from most apical (low frequency) to most basal (high frequency). The band center frequencies are shown on the top *x* axis. ACE, advanced combination encoder; SCE, spectral contrast enhancement.

processor. The playback setup was placed inside a soundproof anechoic chamber, where the signal was picked up by the Cochlear microphone and transmitted to the xPC system which generated the real-time electrical stimulation patterns delivered to the participants.

Procedure

All experiments were conducted within two sessions of maximum 3 hr each, including breaks. Some participants requested to have both sessions conducted on the same day, with a 1- to 1.5-hr break in between the sessions. This was requested by some of the participants who traveled a long distance. Otherwise, each session was conducted on a separate day so as not to exhaust the participants. Experiment 3 was usually conducted in the first session, while experiments 1 and 2 were conducted in the second session.

The participant's clinical map was first loaded to Simulink to obtain their threshold (T) and comfort (C) stimulation levels, their number of maxima, clinical stimulation rate, and frequency-to-electrode allocation map. The control strategy was ACE (SCE factor = 0) with eight maxima.

Next, the loudness balancing procedure deployed by Nogueira et al. (2016) was performed to equate the perceived loudness level of ACE to SCE as follows. A volume setting is implemented in ACE (and subsequently SCE) which allows controlling the stimulation level. This is performed by adjusting a proportion of the participant's dynamic range which is the range between the T- and C-levels (see Nogueira et al. 2016 for more details). The loudness balancing stimulus consisted of presenting a single sentence in a loop. This sentence was chosen from the corpus deployed in experiment 1 and was not used in subsequent data collection. The sentence was calibrated to 65 dB SPL. The volume setting applied to this stimulus in Simulink was set such that the sentence was not perceived by the participant to be too loud or too soft. A loudness scale sheet, identical to the one used in the clinic for fitting purposes, was used to assess the perceived level of the stimulus and ranges between 0 (nothing heard) to 10 (too loud). A comfortable loudness level of 7 (loud enough but pleasant) was selected such that the sentence was loud enough to be intelligible but not eliciting an uncomfortable sensation. The volume setting for the ACE strategy was adjusted by the experimenter in Simulink until the participant reported a perceived loudness level of 7. Next, the experimenter switched the strategy to SCE and asked the participant to rank the perceived loudness of the sentence. The experimenter also adjusted the gain for SCE until the participant reported a loudness level of 7. This loudness balancing procedure was repeated twice, starting at 30% below and above the volume selected for ACE in the first loudness-balancing attempt. The average volume setting for all three trials was then applied to SCE and the participant was asked, as a final check, to judge whether the sentence played through SCE and ACE were of the same loudness. This final check usually indicated that both strategies were at the same perceived loudness level. The average volume setting was then applied to the SCE output for all experiments.

After loudness balancing, a short training block was administered for each experiment, with feedback. Finally, participants were asked to perform the actual test after the training block and were not provided with feedback during data collection, except in experiment 3.

All participants were given both oral and written instructions that appeared on a computer screen placed in front of them. Participants either responded verbally (experiment 1), via a button box (experiment 2), or via a response button that was displayed on the screen (experiment 3).

EXPERIMENT 1: EFFECT OF SCE ON SPEECH-ON-SPEECH INTELLIGIBILITY

Methods

Stimuli • Stimuli were taken from the German HSM sentence test (Hochmair-Desoyer et al. 1997), which is composed of 30 lists of 20 sentences each taken from everyday speech, including questions. Sentences in this corpus are made up of three to eight words, and a single list contains 106 words in total. The original HSM material was recorded from an adult native German male speaker; however, for the purpose of this study, recordings from an adult female speaker were used instead, which were previously recorded at the MHH. The adult native German female speaker had an average F0 of about 218 Hz, and an estimated average VTL of around 14 cm. Because no demographic information about the height of the female speaker was documented, the speaker's height was estimated as 166 cm as explained previously in the section Voice Cue Manipulations. Data on the speaker's height are important because height was shown in the literature to be strongly correlated with VTL (Fitch & Giedd 1999), and thus when the speaker's VTL is unknown, the speaker's height can be used to obtain a good estimate of VTL. These recordings were used because the research questions in this study investigate voice differences starting from a female speaker and moving toward a male-like voice (Fig. 1) to be comparable to the manipulations performed by El Boghdady et al. (2019). Lists 1 to 12 were used in this experiment and were all equalized in root-mean-square intensity.

Target sentences were taken from lists 1 to 8, with one list per experimental condition, masker sentences were taken from lists 9 and 10, and training sentences were taken from lists 11 and 12. For each participant, the list assigned to a given experimental condition was randomly selected without replacement from the target sentence lists. Four different masking voices were created using STRAIGHT according to the parameters shown in Figure 1: the same talker as the target female [resynthesized with $\Delta F0 = 0$ st, $\Delta VTL = 0$ st], a talker with a lower F0 relative to the target female [$\Delta F0 = -12$ st, $\Delta VTL = 0$ st], a talker with a longer VTL relative to the target female [$\Delta F0 = 0$ st, $\Delta VTL = +3.8$ st], and a talker with both a lower F0 and a longer VTL relative to the target female to obtain a male-like voice [$\Delta F0 = -12$ st, $\Delta VTL = +3.8$ st]. These conditions are referred to as *Same Talker*, *F0*, *VTL*, and *F0+VTL*, respectively, in the rest of this manuscript. All target sentences were always kept as the original female speaker from the corpus and were not processed with STRAIGHT.

The parameter values for F0 and VTL were chosen based on the findings of an earlier study, in which CI users were found to exhibit a decrement in SoS intelligibility and comprehension when the voice of the masker was manipulated with F0 and VTL values taken from the top-right quadrant as shown in Figure 1 (El Boghdady et al. 2019). These masker voice manipulations, especially for VTL, were shown to degrade SoS intelligibility compared with the Same Talker condition. Stimulation patterns for these stimuli demonstrated that shortening the masker's VTL (along the top-right quadrant in Fig. 1), which stretches the masker's spectrum along higher frequencies, introduced additional masking to the components of the target signal. The

authors reasoned that this manipulation may have contributed to the detrimental effect on SoS intelligibility observed for the CI group. On the basis of these findings, the authors reasoned that masking voices taken from the lower-left quadrant, as done in the present study, should be expected to yield a benefit in SoS performance for CI users. This is because elongating the masker's VTL is expected to cause the maskers' spectrum to be compressed toward lower frequency components, thereby providing some release from masking. This premise was statistically tested in the Results section of this experiment.

In a given trial, the masker sequence was designed to start 500 msec before the onset of the target sentence and end 250 msec after the offset of the target. The masker sequence was constructed by randomly selecting 1-sec long segments from the masker sentences previously processed with STRAIGHT for the given ΔF0 and ΔVTL condition. A raised cosine ramp of 2 msec was applied both to the beginning and end of each segment, and all segments were concatenated to form the masker sequence. Finally, a 50-msec raised cosine ramp was applied both to the beginning and end of the entire masker sequence.

Target sentences were all calibrated at 65 dB SPL (same level as that used in the loudness balancing procedure), and the intensity of the masker sequence was adjusted relative to that of the target to obtain the required target-to-masker ratio (TMR). To be able to observe variations in intelligibility, the TMR must be chosen in a way that gives performance levels far enough from floor and ceiling. The TMR in this experiment was set to +10 dB based on data from the study by Hochmair-Desoyer et al. (1997), which demonstrated a speech-in-noise intelligibility score in the mid-range of the psychometric function for CI users (between 20 and 60%) for the same material. This was also confirmed with pilot measurements from CI users for the present SoS task.

Objective Evaluation for SCE Factor Selection for the Speech-on-Speech Task • The aim of this objective evaluation was to select an appropriate value for the SCE factor to be used in this study because it was not feasible to test multiple SCE factors given the time constraints of the study. The SCE factors were evaluated in terms of the resulting improvement in simulated TMR across the entire HSM sentences used in this experiment, similar to what was performed in the study by Nogueira et al. (2016). The hypothesis was that an improvement in TMR observed in the simulations could be related to a benefit from SCE relative to ACE in the psychophysics test.

The SCE factors chosen in this simulation were 0 (ACE strategy), 0.5, 1, and 2. The simulated improvement in TMR was estimated using an off-line MATLAB implementation of SCE and the Nucleus MATLAB Toolbox (NMT v. 4.31) from Cochlear, as was performed by Nogueira et al. (2016). First, the target and masker signals were mixed at a TMR of +10 dB, as in the psychophysics task. This mixture of target plus masker signals was used to compute the weights that should be applied to each spectral envelope of this stimulus. The weights would differ depending on the SCE factor chosen; for SCE factor 0, a weight of 1 was applied, yielding no change in the spectral envelope. Next, these weights were applied to each band of the target and masker signals separately, such that the TMR could be computed. Note that these weights change from frame to frame. However, this technique assumes that the signal processing pathway (Fig. 2) only performs linear operations, which is clearly not the case, as in the envelope extraction block. To circumvent this issue, the weights were applied to the target and masker signals separately and each signal was processed until (and including) the fast Fourier transform block. This procedure was carried out for all sentences in the HSM corpus that were used in the psychophysics test, and all masker voice conditions were also evaluated.

Figure 4 shows the average improvement in simulated TMR for each SCE factor (0.5, 1, and 2) relative to ACE (SCE factor = 0) processing of the Same Talker condition. The plot shows the expected improvement in TMR as a function of the SCE factor, in addition to the expected benefit from the masker voice differences all relative to the Same Talker condition. Error bars denote one standard error (SE) of the mean. Consistent with previous literature, which shows an advantage of SCE for speech-in-noise intelligibility (e.g., Baer et al. 1993; Nogueira et al. 2016), these simulations also reveal that SCE has the potential of providing improvements in TMR compared with ACE for SoS. The simulations demonstrate that benefit in TMR relative to ACE appears to be consistent across the different masker voices, with larger SCE factors expected to yield a larger benefit. In the psychophysics test, only the SCE factor of 0.5 was tested because it has been shown previously to yield a benefit in speech-in-noise intelligibility for CI users compared

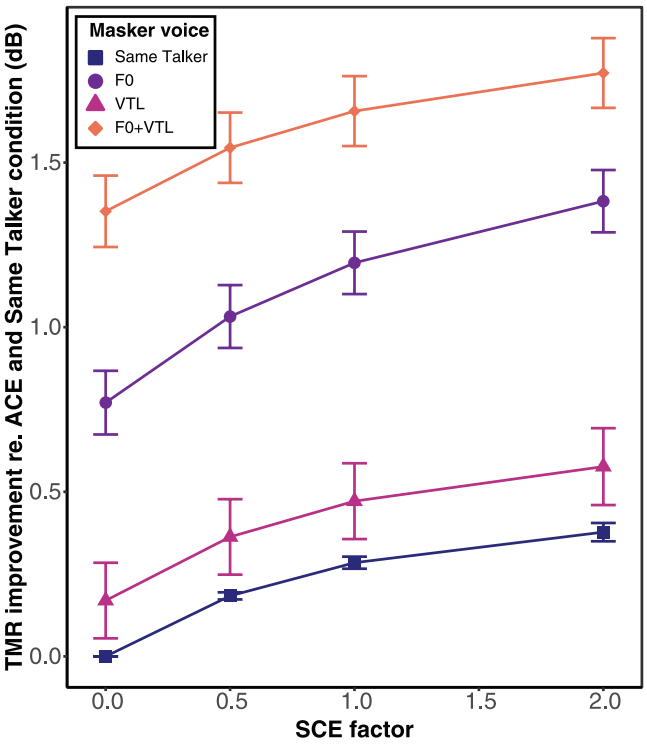

Fig. 4. Improvement in simulated TMR for the different SCE factors (0.5, 1, and 2) relative to ACE (SCE factor = 0) processing of the Same Talker condition. Simulations were obtained using the Nucleus MATLAB Toolbox (NMT, v 4.31) from Cochlear, with an initial input TMR of +10 dB as explained in the text. The different curves represent the masker voice conditions tested in the psychophysics experiment. The error bars indicate the SE of the mean TMR. *Same Talker:* Masker is the same speaker as the target. *F0:* Masker has a lower F0 relative to the target speaker. *VTL:* Masker has a longer VTL relative to the target. *F0+VTL:* Masker has both a lower F0 and a longer VTL relative to the target speaker. ACE, advanced combination encoder; NMT, Nucleus MATLAB Toolbox; SCE, spectral contrast enhancement; VTL, vocal tract length.

to loudness-balanced ACE (Nogueira et al. 2016), while no significant difference was observed between the performance under SCE factor 0.5 and SCE factor 1 for stimuli that were not loudness balanced. In addition, Baer et al. (1993) suggested that severe SCE may in fact lead to worse speech-in-noise intelligibility. However, future testing of additional SCE factors may be beneficial to further assess the effects of SCE processing.

The simulations also demonstrate an expected benefit from masker voice differences compared to the Same Talker condition, with conditions F0 and F0+VTL yielding the largest expected benefit. Contrary to what was expected, the effect of SCE did not differ depending on the masker voice. The expectation was that the effect of SCE would be larger for voice manipulations along the VTL dimension because of its largely spectral nature. These observations are compared with the psychophysical data in the Results section of experiment 1.

Procedure • The SoS paradigm for this experiment was based on that used in the study by El Boghdady et al. (2019). A single target-masker combination was presented per trial to a participant, and the participant was asked to concentrate on the target sentence and attempt to ignore the masker. Participants were asked to repeat whatever they thought they heard from the target sentence, even if it was a single word, a part of a word, or if they thought what they heard did not make sense.

Trials were blocked per strategy, meaning that a participant would perform all conditions for a given strategy before switching to the second one. This was done to prevent the extra time needed for switching back and forth between strategies at the beginning of each condition, as the strategy selection was manually performed in Simulink. The starting strategy (ACE or SCE) was randomized and counterbalanced across participants, such that seven participants started with ACE, while the other six started with SCE. Participants were blinded to the strategies tested.

At the beginning of each strategy block, short training was provided to familiarize the participants with the sound of the strategy tested before actual data collection. The training for the first strategy tested was always assigned 12 sentences randomly selected from list 11, while the training for the second strategy was assigned 12 sentences from list 12. The training for a given strategy was divided into two parts. In the first part, six sentences were presented in quiet to accustom the participants to the voice of the target female speaker. In the second part, the remaining six sentences were presented in the presence of a masker speaker at a TMR of +14 dB to acquaint the participants with the SoS task itself. This masker had a combination of ΔF0 and ΔVTL of −6 st and +6 st, respectively, which, while also falling in the bottom-left quadrant of Figure 1, were still different from those used during data collection. This was carried out so as not to bias participants toward a particular voice condition that would be used during data collection. During the entire training (quiet and SoS), both auditory and visual feedback were provided after the participant's response, such that the target sentence was displayed on the screen while the entire stimulus was replayed once more through the loudspeaker.

Data collection was composed of a total of 160 trials for both strategy blocks together (20 sentences per list × 4 voice conditions × 2 strategies), which were all generated off-line before the experiment began. The trials within a strategy block were all pseudo-randomized. No feedback was provided during data collection: participants only heard the stimulus once and were not shown the target sentence on the screen.

The verbal responses were scored online by the experimenter, on a word-by-word basis, using a graphical user interface (GUI) programmed in MATLAB. For each correctly repeated word, the experimenter would click the corresponding button on the GUI which was not visible to the participant. In addition, the verbal responses were recorded and stored as data files to allow for later off-line inspection. A second GUI was programmed to allow the experimenter to listen to the responses off-line and double-check if there were incorrectly scored words during the online procedure.

Response words were scored according to the following guidelines. The HSM sentences contain words that are hyphenated in the corpus, such as "Wochen-ende." These words in German are written without the hyphen but are hyphenated in the HSM corpus to enable scoring each part of the word separately. If the participant repeated a part of such words, only that part was marked as correct, while if they repeated both parts correctly, both parts were marked as correct. This was slightly different from the scoring paradigm followed in the study by El Boghdady et al (2019); however, the Dutch corpus used in the latter study does not include such hyphenated words. In addition, no penalty was given if a participant changed the order of the words in the sentence or added extra words.

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." In addition, confusion of adjective form, for example, saying "keiner" instead of "keine," or confusing the Dativ with the Akkusativ article, for example, confusing "der" with "dem" or "den," was also considered incorrect. Confusion of verb tenses or incorrect verb conjugation was considered incorrect. In addition, if the participant only uttered a single pronoun, like "he," "she," or "I," it was considered incorrect even if it was in the sentence, as this might have constituted a guess.

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. In addition, the experimenter could also ask the participant to take a break if they judged it to be necessary.

Statistical Analyses

All data analyses were performed in R (version 3.3.3, R Foundation for Statistical Computing, Vienna, Austria, R Core Team 2017), and linear modeling was done using the *lme4* package (version 1.1-15, Bates et al. 2015). To quantify the main effect of strategy and voice on SoS intelligibility, an analysis of variance (ANOVA) was applied to a logistic regression model, as defined by Equation 1 in *lme4* syntax. The Chi-squared statistic (χ^2) with its degrees of freedom and corresponding *p* value are reported from the ANOVA.

$$score \sim strategy * voice + (1 + strategy * voice | participant). \quad (1)$$

In Equation 1, *score* denotes the per-word score (0 or 1) as the predicted variable and the term *strategy* * *voice* denotes the fixed-effects of strategy (ACE versus SCE), masker voice, and their interaction. Interaction effects give insight into whether a fixed effect is consistent across the levels of other factors. The terms inside the parentheses denote the random effects estimated per participant, such that, for each participant, a random intercept ("1+" term inside the parentheses) and a random slope

for each of *strategy*, *voice*, and their interaction are estimated (*strategy* * *voice* | *participant* term). The random effects defined by this model assume a different baseline performance for each participant in addition to a different benefit from SCE and masker voice condition relative to the baseline performance. The estimated coefficients for each fixed factor of the model (β), the associated SE, Wald's *z* value, and corresponding *p* value are reported.

In addition, to characterize the effect of strategy for each masker voice (post hoc analyses), a separate logistic regression model, as defined by Equation 2, was applied for each masker voice condition. A false-discovery rate correction (Benjamini & Hochberg 1995) was then applied to all *p* values obtained from these per-voice-condition models to correct for multiple comparisons.

$$score \sim strategy + (1 + strategy \mid participant). \quad (2)$$

Results and Discussion

Figure 5 shows the distribution of SoS intelligibility scores for each masker voice condition under each strategy. The figure demonstrates an overall benefit in SoS intelligibility scores for the SCE strategy compared to ACE. The overall main effect of strategy in addition to the effect of strategy for each voice condition is reported in detail later.

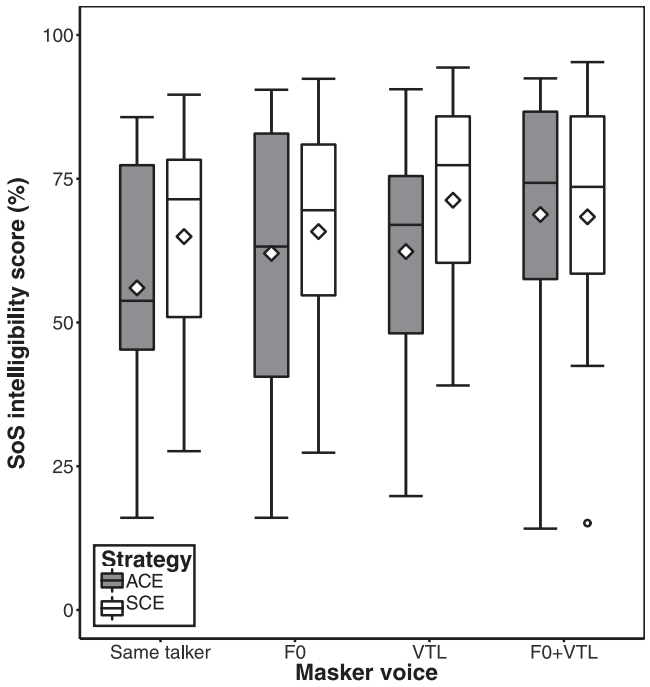

Fig. 5. Distributions of SoS intelligibility scores across participants for each masker voice condition under ACE (gray boxes) and SCE factor = 0.5 (white boxes). *Same Talker:* the condition when the target and masker were the same female speaker [ΔF0 = 0 st, ΔVTL = 0 st]. *F0:* the condition when the masker had a lower F0 [ΔF0 = −12 st, ΔVTL = 0 st] relative to that of the target speaker. *VTL:* the condition when the masker had a longer VTL [ΔF0 = 0 st, ΔVTL = +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 [ΔF0 = −12 st, ΔVTL = +3.8 st] relative to those of the target (see Fig. 1). 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 interquartile range (IQR). Diamond-shaped symbols denote the mean SoS intelligibility score, while circles indicate individual data outside of 1.5 times IQR. ACE, advanced combination encoder; VTL, vocal tract length.

General Effect of Masker Voice and Strategy • The ANOVA described in the Statistical Analyses section revealed an overall significant main effect of strategy [$\chi^2(1) = 12.45$, $p < 0.001$], masker voice [$\chi^2(3) = 38.07$, $p < 0.001$], and their interaction [$\chi^2(3) = 15.93$, $p = 0.001$] on SoS intelligibility scores. The significant effect of the interaction between strategy and masker voice indicates that either the direction of the effect of strategy or its magnitude differs depending on the type of masker voice.

Post hoc Analyses: Benefit from Changing Masker Voice Compared to Same Talker Condition • The coefficients from the logistic regression model reveal a more detailed picture of the nature of the effects observed from the ANOVA. One of the goals of this analysis was to check whether the chosen F0 and VTL manipulations indeed yielded a benefit in SoS intelligibility for the CI users tested here. In this logistic regression, the SoS intelligibility scores under ACE processing (all gray boxes as shown in Fig. 5) for all F0 and VTL manipulations were compared to those obtained for the baseline condition when the target and masker were the same female speaker (*Same Talker* condition). Compared to the *Same Talker* condition, SoS intelligibility scores were found to improve when the masker's VTL ($\beta = 0.31$, SE = 0.15, $z = 2.00$, $p = 0.046$) or both the masker's F0 and VTL were different from those of the target ($\beta = 0.70$, SE = 0.14, $z = 4.98$, $p < 0.001$). However, no difference in SoS intelligibility was observed between the *Same Talker* and *F0* conditions ($\beta = 0.33$, SE = 0.19, $z = 1.80$, $p = 0.07$). This indicates that, as hypothesized, the male voice space tested here indeed provided a benefit in SoS intelligibility from voice differences between target and masker.

Results from the simulations carried out at the beginning of this manuscript do not seem to agree with what has been reported in the psychophysics. The simulations show that the voice benefit for CI users was expected to be largest for masker conditions F0 and F0+VTL. This is consistent with data in the literature showing that CI users have more usable F0 cues than VTL cues for tasks such as gender categorization (e.g., Fuller et al. 2014), in addition to their reasonable sensitivity to F0 differences compared to VTL differences (Gaudrain & Başkent 2018). Thus, it may be expected that CI users should benefit more from F0 differences compared to VTL differences in SoS situations, as shown in the simulation data. In contrast, the psychoacoustic data revealed no benefit from masker condition F0, and masker condition VTL was found to provide a significant benefit in SoS intelligibility. This discrepancy between simulation and real data might be due to the fact that the simulations only consider aspects from the signal processing side (energetic masking), but do not account for any models of perception (perceptual effects of both energetic and informational masking).

Consistent with the first hypothesis, the overall logistic regression model also revealed a significant benefit in SoS intelligibility from SCE processing compared to that of ACE for the baseline condition when the masker and target were the same talker ($\beta = 0.43$, SE = 0.17, $z = 2.56$, $p = 0.01$). The logistic model coefficients also revealed that the effect of SCE was consistent for all masker voice conditions except F0+VTL, as indicated by the significant interaction term in the logistic regression model ($\beta = -0.49$, SE = 0.18, $z = -2.75$, $p = 0.006$).

To test for the effect of SCE under each masker voice condition separately, the following analyses were performed.

TABLE 2. Coefficients for the Effect of Strategy Obtained from the Logistic Regression Model Applied Separately for Each Voice Condition

Masker Voice Condition	Strategy
Same talker	β = 0.43, SE = 0.17, z = 2.53, p = 0.02*
F0	β = 0.17, SE = 0.15, z = 1.13, p = 0.34
VTL	β = 0.47, SE = 0.12, z = 4.05, p < 0.001***
F0+VTL	β = -0.04, SE = 0.10, z = -0.45, p = 0.66

β represents the estimated parameter from the logistic regression, SE is the standard error of that estimate, z is the Wald-z statistic, and p is the p value after FDR correction for multiple comparisons.
**p < 0.05; ***p < 0.001.*
FDR, false-discovery rate; VTL, vocal tract length.

Post hoc Analyses: Effect of Strategy for Each Voice Condition • A separate logistic regression representing score as a function of strategy was applied to each voice condition separately, following the model in Equation 2. The results are provided in Table 2. This analysis revealed that SCE provided a benefit in SoS intelligibility in two out of four masker voice conditions; namely, in the Same Talker and VTL conditions. For masker voice conditions F0 and F0+VTL, no difference in SoS intelligibility scores between SCE and ACE was observed. These observations indicate that the effect of SCE becomes small when a difference in F0 is introduced between target and masker speakers.

For masker condition *F0*, the lack of effect of SCE may be attributed to the individual variability across the CI users tested. For example, in the individual data as shown in Figure 6, in which almost half of the participants (P04, P05, P06, P08, P11, and P12) exhibit a benefit from SCE under masker condition F0, while the other half of the participants either do not show any effect or demonstrate a decrement in performance. This indicates that the effect of SCE on SoS intelligibility when talkers differ in F0 may be subject-dependent. However, another explanation for this effect could be that SCE does not improve performance for CI listeners as a group under this specific masker condition.

For masker condition *F0+VTL*, a possible explanation for not observing a benefit from SCE could be that participants already gained a large enough benefit from that particular voice difference relative to the *Same Talker* condition. This means that any additional improvements from SCE processing may have been masked by the voice benefit from the F0+VTL difference, as can be observed as shown in Figure 5. Taken together, the results from this experiment demonstrate that SCE has the potential to improve intelligibility of individual words under adverse masking conditions, especially if the masker has voice characteristics that are similar to those of the target, or if the difference between the two competing talkers lies along VTL.

In the next experiment, the effect of SCE on overall sentence comprehension in the presence of a competing talker was assessed. According to the study by Kiessling et al. (2003), "Comprehending is an activity undertaken beyond the processes of hearing and listening [and] is the perception of information, meaning or intent." CI processing inevitably leads to some distortions in the acoustic signal and can thus impair overall sentence comprehension if a sufficient number of words is distorted beyond the ability of top-down reconstruction by the brain. In the following experiment, the effect of SCE on SoS comprehension was evaluated because it more closely captures realistic listening situations (Best et al. 2016) in which a listener assigns meaning to an entire auditory stream (Rana et al. 2017). SoS comprehension accuracy and speed (reaction times; RTs) were measured using a sentence verification task (SVT), as was performed by Baer et al. (1993). Baer et al. have demonstrated that RTs measured from SVT were able to capture potential benefits of SCE processing. Thus, in the context of this study, RT measures could reveal an effect of SCE for SoS comprehension.

The literature has argued that interpreting accuracy and RT measures in isolation of one another may be challenging, because a participant may trade speed for accuracy (e.g., Schouten & Bekker 1967; Pachella 1974; Wickelgren 1977). This speed-accuracy trade-off can be addressed by combining accuracy and RT measures into a unified measure of performance called the *drift rate* (for a review, see Ratcliff et al. 2016), which quantifies the quality of information accumulated by the CI listener until they give a response. The drift rate was computed using the EZ-diffusion model (Wagenmakers et al. 2007) which is a simplified version of the full model proposed by Ratcliff (1978). The EZ-diffusion model utilizes the proportion of correct responses and the RT distribution (mean and variance) of the correct responses to compute the drift rate.

EXPERIMENT 2: EFFECT OF SCE ON SPEECH-ON-SPEECH COMPREHENSION

Methods

Stimuli • The voice conditions for the masker speaker in this experiment were the same as those in experiment 1. The masker sequence was also 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 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 • Translation from Dutch to German was performed by three independent native German speakers: two of those speakers were also fluent in Dutch, while the third had sufficient knowledge of the language. The three translated versions were consolidated together to give the least ambiguous structures and then were relayed to a fourth translator for a blinded back translation from German to Dutch. This translator was a native Dutch speaker who was also fluent in German and had not been exposed to the original Dutch sentences. The back translations were then checked against the original Dutch version for consistency. One sentence pair lost its meaning when

*Contrary to the English SVT developed by Baddeley et al. (1995), the Dutch SVT developed by Adank and Janse (2009) is not divided into lists. These corpora also slightly differ from the SVT developed by Pisoni et al. (1987), such that the resolving word, which determines whether the statement is true or false, is not always at the end of the sentence. This could potentially influence response time measurements since such measurements are usually marked starting from the offset of the resolving word. This issue has been addressed while analyzing the reaction time data.

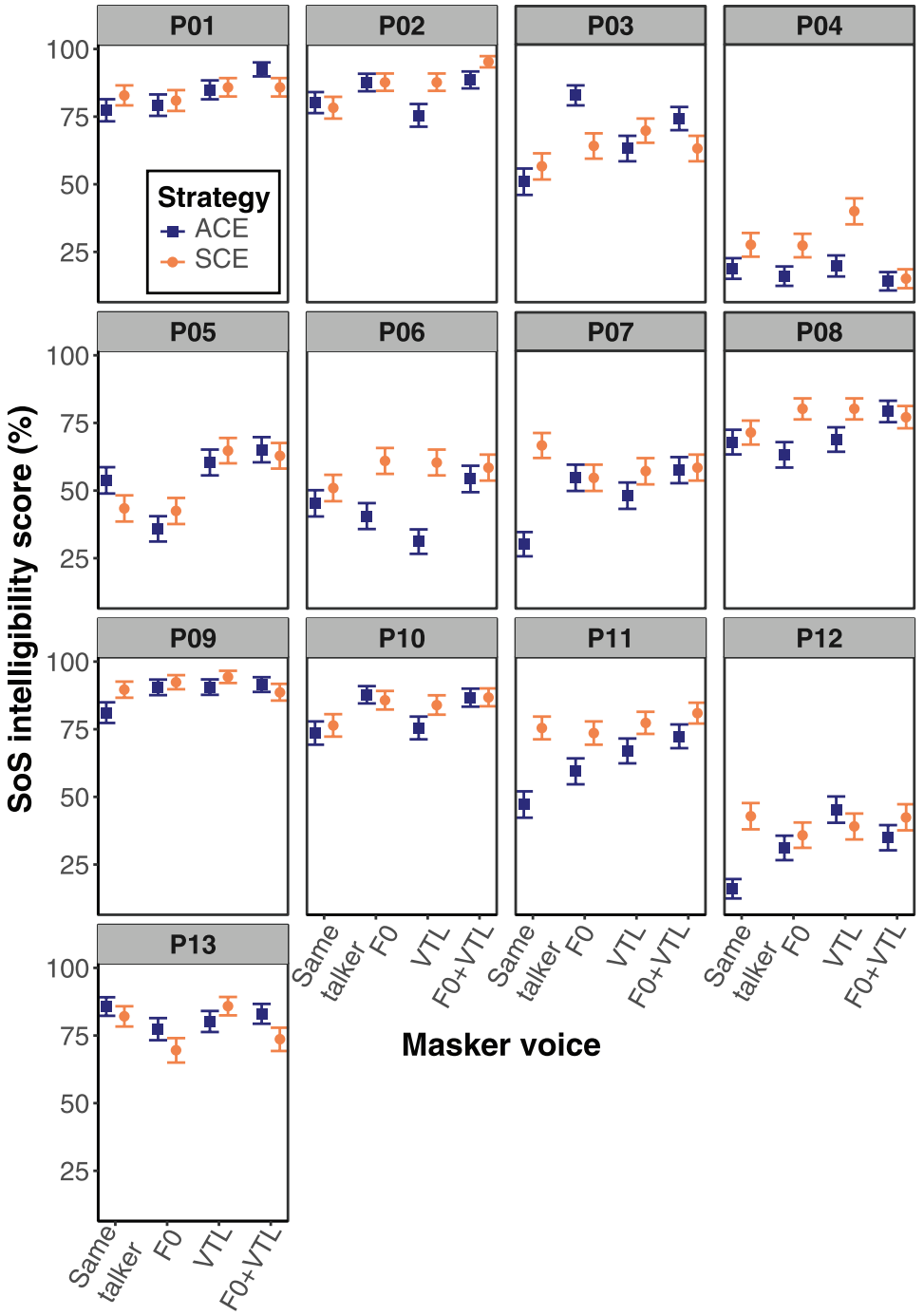

Fig. 6. Individual data for the mean SoS intelligibility scores for each masker voice condition (see Fig. 1). Dark squares denote intelligibility scores obtained with the ACE strategy, while bright circles indicate scores obtained using SCE. The error bars denote one SE of the mean. ACE, advanced combination encoder; SCE, spectral contrast enhancement.

translated to German and was thus discarded from the translations, resulting in a total of 99 true-false sentence pairs in the German corpus. The additional four-sentence pairs introduced by El Boghdady et al. (2019) for training purposes were also translated to German. This was done to ensure a sufficient number of training and test sentences. Appendix A (Supplemental Digital Content 1, http://links.lww.com/EANDH/A696) in the supplementary material provides the true and false sentence pairs for both the Dutch and German versions.

Recordings and Processing • Recordings were made in a sound-proof anechoic chamber at the University Medical Center Groningen, NL, using an RØDE NT1-A microphone mounted on an RØDE SM6 with a pop-shield (RØDE Microphones LLC, California, USA). The microphone was connected to a PreSonus TubePre v2 amplifier (PreSonus Audio Electronics, Inc., Los Angeles, USA) set at an amplification of 10 dB, with 80 Hz noise cancellation and phantom power activated. 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.

Recordings were taken from a 27-year-old native German female speaker from Falkenstein/Harz, with an average F0 of 180 Hz, and an estimated VTL of about 14.1 cm. Her VTL was estimated based on her height of 167 cm following the method provided by Ives et al. (2005) and the data from Fitch and Giedd (1999).

The speaker was instructed to stand on a cross on the ground of the testing booth marking a distance of about 1 m from the microphone. An additional set of recordings was made for lists 9 and 10 from the HSM corpus, which was used to construct the maskers.

Sentences were presented in three rounds to the speaker in a slideshow on a touchscreen inside the soundproof booth, in which the sentence order differed per round. The speaker was instructed to read the sentence twice silently, and then articulate it as clearly as possible and at a fixed rate with a neutral voice tone. This procedure yielded three recordings per sentence, from which the recording with clearest articulation was chosen.

Individual sentences were then manually extracted from the recordings. Important articulation cues at the onset and offset of the sentence were maintained by including a minimum duration of 20 msec before the onset and 50 msec after the offset of the sentence, and each sentence file did not exceed 3.5 sec. Cosine ramps of 50 msec and 100 msec were applied at the beginning and the end, respectively, to minimize sudden onset and offset effects, respectively. All sentence files were then equalized in root-mean-square intensity. Pilot tests with young NH native Dutch and native German speakers revealed no significant differences in either accuracy scores or RT distributions between the Dutch and German corpora.

Procedure • Following the paradigm of Adank and Janse (2009) and Pals et al. (2020) for the SVT, participants were instructed to indicate whether the target sentence was true (labeled as WAHR) or false (labeled as UNWAHR) by pressing the corresponding button on a button-box as quickly and accurately as possible within a specific time window. In the present experiment, a longer time window (6 sec) than that used in the aforementioned studies was used so as not to stress the CI users during testing. If participants did not respond within that time window, the response was recorded as a *no-response*, and the experiment proceeded to present the next stimulus. Participants were also instructed to provide the first response that occurred to mind without overthinking. Participants were allowed to respond at any time during stimulus delivery, similar to the procedure carried out by Adank and Janse (2009) and Pals et al. (2020), to allow measuring RTs relative to the end of the resolving word (see Footnote*). This could potentially result in negative RTs if the participant gave a response before the offset of the resolving word.

As was done in experiment 1, trials here were also blocked per strategy, with the starting strategy randomized and counterbalanced across participants. At the beginning of each strategy block, a short training was provided to acquaint the participants both with the task and the strategy. The last eight sentence pairs in Appendix A (Supplemental Digital Content 1, http://links.lww.com/EANDH/A696) were assigned to training and were not used in actual data collection. Out of these eight pairs (16 true-false sentences), four true and four false sentences were randomly picked and assigned to the training block of the first strategy tested, while the remaining four true and four false sentences were assigned to the training of the second strategy tested. No true-false pair was assigned to the same training block.

Each training block was split into two parts. In the first part, the training sentences were presented without a competing masker to accustom the participants to the sound of the target speaker's voice through the tested strategy. In the second part of the training block, a competing masker was added with the same voice parameters as those of the training masker voice used in experiment 1 but at a training TMR of +14 dB. Both audio and visual feedback were provided for both parts of the training (quiet and SoS) as was done in experiment 1: participants were shown whether the sentence was true or false and the sentence was also shown on the screen while the whole stimulus was replayed through the loudspeaker.

The remaining sentences that were not used in training were used for data collection. These sentences were distributed among the number of tested conditions (4 masker voice conditions × 2 strategies), and sentences of a true-false pair were never assigned to the same condition. The input TMR was the same as that used in experiment 1 (+10 dB). All stimuli were generated off-line for both strategy blocks and pseudo-randomized within each block. During data collection, no feedback was given to the participants. The entire experiment lasted for a maximum of 1 hr, including breaks.

Statistical Analyses • SoS comprehension accuracy scores were converted to the sensitivity measure *d'* (Green & Swets 1966) because this measure is unbiased to a participant's preference for a particular response. Both the *d'* and drift rate data were analyzed using a linear mixed-effects model (*lmer* function in R), with the same parameters as in Equation 1, but without random slope estimates for the interaction effect per participant to improve model convergence. RT data were analyzed using a generalized linear mixed-effects model (*glmer* function in R) with the same parameters as shown in Equation 1.

Only RTs to correct responses were analyzed. Because participants could respond at any time during stimulus delivery, negative RTs were possible, although their occurrence was rare (amounted to 0.47% of the analyzed RT data). They were thus discarded to allow fitting the positively skewed RT distribution to an inverse Gaussian distribution following the recommendations provided by Lo and Andrews (2015) for analyzing RT data. The resulting model for RTs was of the form $\frac{-1}{RT} = \beta_0 + \beta_{strategy} + \beta_{masker\ voice} + \beta_{strategy:masker\ voice}$, where RT is expressed in seconds, and where $\beta_{strategy}$ takes different values for every strategy, $\beta_{masker\ voice}$ takes different values for every masker voice, and $\beta_{strategy:masker\ voice}$ takes different values for every combination of strategy and masker voice. To determine the overall main effect of strategy and masker voice on each of the three aforementioned performance measures (*d'*, RTs, and drift rate), an ANOVA was applied to the linear regression models as was done in the previous experiment.

Results and Discussion

Figure 7 shows the SoS comprehension accuracy (in *d'*), RTs (in seconds), and drift rates (in arbitrary units per second) as a function of processing strategy for each type of masker voice (individual data provided in Appendix B, Supplemental Digital Content 1, http://links.lww.com/EANDH/A697). While

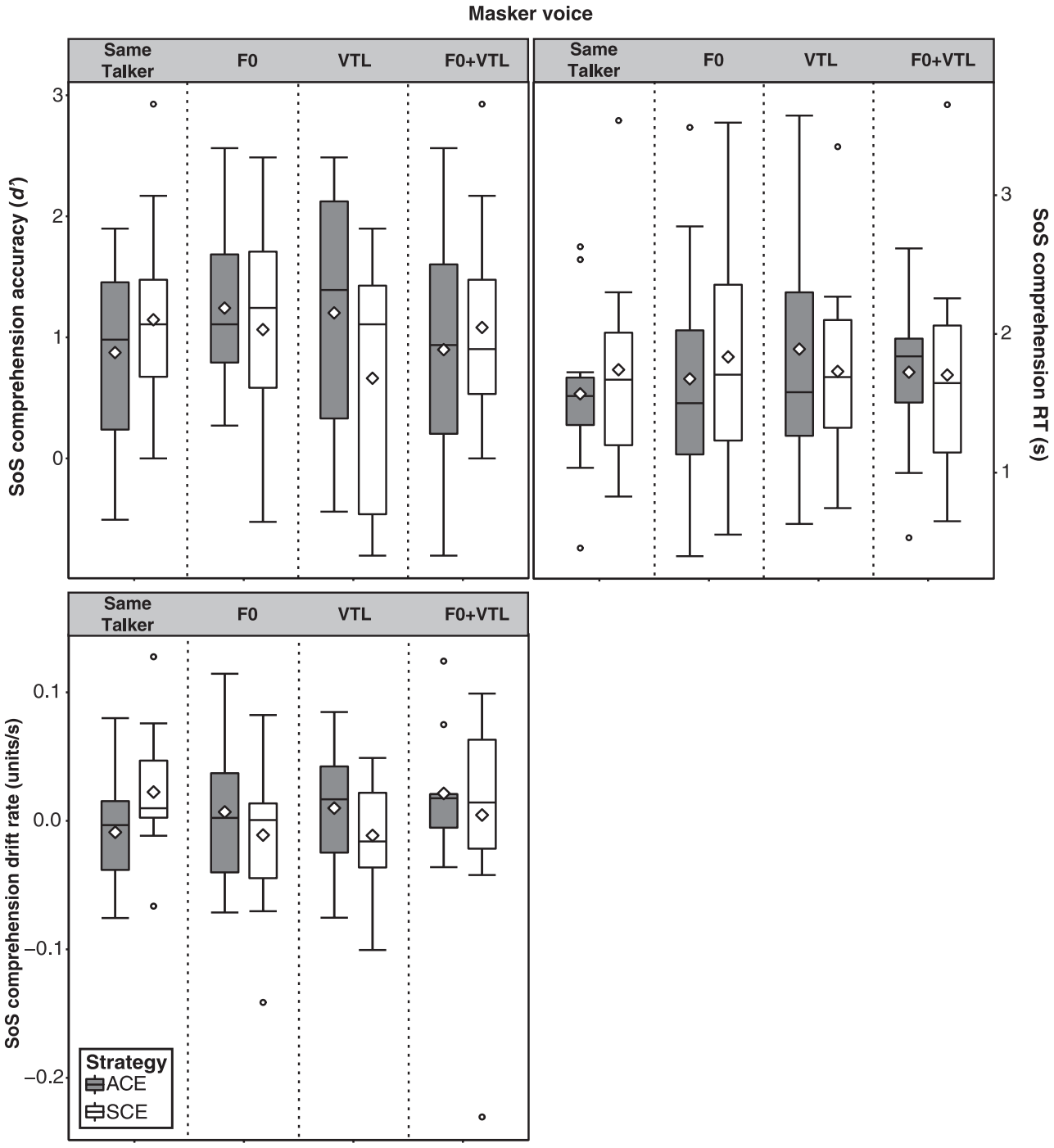

Fig. 7. SoS comprehension accuracy in *d'* (top left), RTs (top right), and drift rates (bottom left) for the different masker voices tested under ACE (gray bars) and SCE processing (white bars). Boxplot statistics are the same as those described in the caption of Figure 5. ACE, advanced combination encoder; RT, reaction time; SCE, spectral contrast enhancement.

previous studies have demonstrated that RTs can reflect differences in listening conditions (e.g., Gatehouse & Gordon 1990; Baer et al. 1993; Adank & Janse 2009; Pals et al. 2015), the data from this experiment did not reveal marked differences in SoS comprehension performance between ACE and SCE. The statistical analyses confirmed these observations: no effect of strategy or masker voice was observed for either the *d'*, RT, or drift rate data ($p > 0.06$ for all main effects). The data from this task do not support the hypothesis that SCE processing could improve SoS comprehension for CI users.

It is important to note that sentences in the SVT material were much shorter compared to those from the HSM corpus. Thus, if participants missed the first or last words in an SVT sentence, they were more likely to get an incorrect response because they were not able to collect enough information to make a valid judgment about the truth of the sentence. Moreover, some participants verbally reported that they found this

task to be more difficult than that administered in experiment 1, and thus some participants either completely refrained from performing the SVT or could not respond within the dictated time window to entire conditions. Discarding the data from those participants and repeating the statistical analyses did not influence the pattern of results obtained.

On the basis of the findings of both experiments, because SCE was already observed to yield a small yet significant improvement in SoS intelligibility scores, especially for some CI users in a consistent manner, the question of whether this improvement stems from an improvement in the sensitivity to voice cue differences (F0 and VTL) between target and masker speakers was investigated in the following experiment.

EXPERIMENT 3: EFFECT OF SCE PROCESSING ON SENSITIVITY TO F0 AND VTL CUES

Methods

Stimuli • Because the CI users recruited were native German speakers, the JND task deployed in previous studies (e.g., Gaudrain & Başkent 2015; El Boghdady et al. 2018; Gaudrain & Başkent 2018; El Boghdady et al. 2019) for measuring JNDs was adapted to German in the following manner. In the aforementioned studies, the stimuli were taken from the Dutch Nederlandse Vereniging voor Audiologie corpus (Bosman & Smoorenburg 1995), which is a test typically used in the clinic to measure phoneme recognition in quiet, and is comprised of common monosyllabic Dutch words (e.g., "Bos," "Vaak," and "Boom"). In the present study, stimuli were taken from the Freiburg monosyllabic word test (Hahlbrock 1953), which is a German test used in the clinic to test word recognition in quiet. This test consists of common monosyllabic German words (e.g., "Bach," "Nuß," and "Zahl"), and thus was considered a good equivalent to the Dutch corpus for the purpose of this test. The German Freiburg words were recorded in this study from a 29-year-old native German female speaker from Wesel, with an estimated average F0 of 233 Hz and an estimated VTL of 13.9 cm (based on a height of 164 cm and the data from the study by Fitch and Giedd, 1999).

Recordings were made in the same manner and using the same setup as those obtained in experiment 2. Seventy-five consonant-vowel (CV) syllables were manually extracted from the recorded words in the corpus, yielding a list consisting of combinations of the consonants (b, d, f, g, h, k, l, l̩, m, n, p, ʁ, z, ʃ, t, v, x, ts) and vowels (iː, oː, uː, a, ɛ, ɪ, ʊ, ɔ, eː). All extracted syllables were equalized in root-mean-square intensity.

The stimuli in this experiment were triplets of CV syllables. These were created by randomly selecting three different CV syllables from the list of 75 available syllables and concatenating them together, with a 50-msec silence gap in-between, to form a triplet. This selection of syllables was different in each trial. Within a given trial, the same triplet of syllables was presented three times, with a silence gap of 250 msec between each presentation. Only one of the three presentations was processed to have a different voice (lower F0, longer VTL, or both, according to the voice vectors as shown in Fig. 1) relative to the other two identical presentations, whose voice was that of the original female speaker. However, all three presentations were always resynthesized using STRAIGHT, even if no F0 or VTL differences were introduced relative to the original female speaker. Thus, the procedure required participants to select the presentation (triplet) that had a different voice relative to the other two in an adaptive three-interval, three-alternative forced choice task.

Procedure • JNDs in this experiment were measured along three voice vectors as indicated by the direction from the origin of the [ΔF0, ΔVTL] plane to the red crosses as indicated in Figure 1. The JND measurement for each of the three voice vectors was repeated three times per strategy yielding a total of 18 experimental conditions (3 voice vectors × 3 repetitions each × 2 coding strategies).

Experimental conditions were blocked per strategy as was done in the previous two experiments. The starting strategy was randomized and counterbalanced across participants, and the order of conditions within a given strategy block was pseudo-randomly shuffled.

The JND measurement for a given voice vector was obtained using a two-down one-up adaptive procedure, yielding 70.7% correct responses on the psychometric function (Levitt 1971), and consisted of a number of trials with three presentations of the same triplet as explained in Stimuli section. The initial trial started with the target triplet having a difference of 12 st relative to the two identical reference triplets. After two consecutive correct responses, this difference of 12 st was reduced by 2 st, and after a single incorrect response, the difference between the reference and target triplets was increased by the same step size. If the difference between the reference and target triplets became less than twice the step size, the step size was reduced by a factor of $\sqrt{2}$. This procedure terminated after eight reversals and the JND was calculated as the mean of the difference in semitones between the reference and target triplets on the last six reversals. The measurement was automatically discarded if the participant did not manage to reach 8 reversals within a maximum of 150 trials. For discarded measurements, an additional attempt was made to obtain the JND. However, none of the participants tested experienced this issue.

A short training was always administered before the beginning of each strategy block to familiarize the participants both with the procedure and with the strategy. Two voice vectors were used during training: [ΔF0 = +5 st, ΔVTL = −7 st] and [ΔF0 = −12 st, ΔVTL = +3.8 st]. Each training condition was programmed to terminate after only six trials. Visual feedback was always provided during training and data collection.

Statistical Analyses • To quantify the effect of SCE on the overall JNDs, a linear mixed-effects model was applied to the log-transformed JNDs (because the step-size is geometrically adapted and because the data are otherwise not normal as they are only positive), with the same parameters as those in Equation 1. A type III ANOVA was then applied to the model.

Results and Discussion

Raw JNDs • Figure 8 shows the raw JND distributions obtained across participants for each voice vector under each strategy (individual data provided in Appendix B, Supplemental Digital Content 1, http://links.lww.com/EANDH/A697). Lower JNDs denote higher (better) sensitivity to the voice cues and vice-versa. The data did not reveal marked differences in performance between the two strategies along any of the voice vectors tested. This was confirmed by the statistical analyses which revealed no overall effect of SCE compared to ACE [$F(1,13.16) = 1.44$, $p = 0.25$].

The first aim of this experiment was to determine the effect of SCE on the sensitivity to F0 and VTL cues, both in isolation and together. Contradicting the second hypothesis of this study, the psychophysical data revealed no perceptual differences between ACE and SCE in terms of JNDs. Figure 9 shows the difference between ACE and SCE in the peaks selected after the N-of-M block, and their proportion of occurrence across all temporal frames and all stimuli tested. The plot shows that in most temporal frames, there is no difference in peak selection between ACE and SCE, as indicated by the large proportion of occurrence for a peak selection difference of 0 (same stimulated electrodes). However, in some cases, the peak selection in SCE differs from that of ACE, but by only a small number of electrodes. Nevertheless, these differences in stimulation patterns between SCE and ACE might not have been perceptually salient enough to elicit a change in the perceived sensitivity to F0 and VTL differences. This was observed in the psychophysics data, such that there was no improvement in either F0 or VTL JNDs from SCE processing.

Correlation Between Benefit From SCE on SoS Intelligibility and JND Tasks • Figure 10 shows the correlations between the benefit in SoS intelligibility scores and JNDs obtained from SCE relative to ACE for each voice condition. For example, the benefit in SoS intelligibility obtained from SCE under masker voice condition F0 was plotted against the benefit in F0 JNDs obtained from SCE. The benefit in SoS intelligibility on the participant level was obtained from the random slope estimated per participant for the effect of strategy per voice condition, which was computed from the linear regression models in experiments 1 and 3. This means that a positive value for this estimate per participant denotes a benefit (better scores for that participant when using SCE compared to ACE) while a negative value denotes a deficit (better scores for ACE than for SCE). The benefit in JNDs was obtained in the same way but multiplied by a negative sign so that positive values would denote a benefit in JNDs from SCE processing (because smaller is better for JNDs).

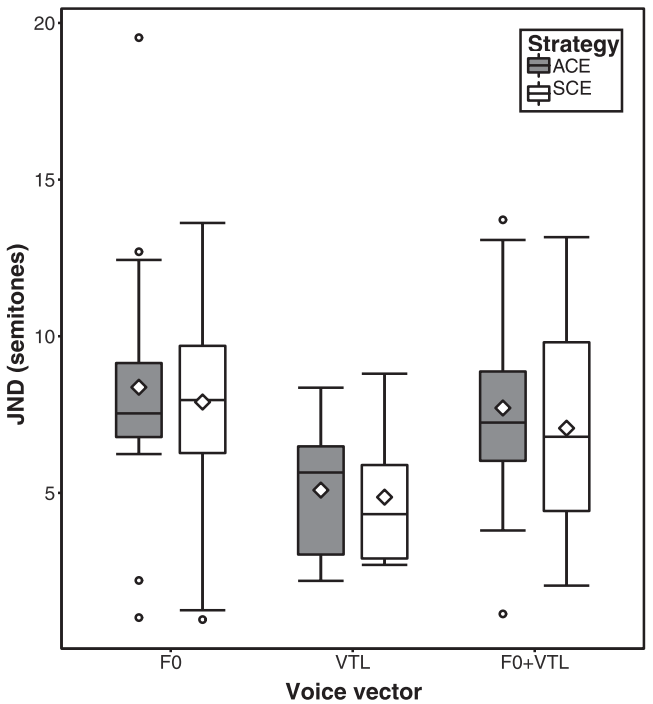

Fig. 8. JND distributions obtained for each voice vector [along negative ΔF0, along positive ΔVTL, and along the diagonal with the combination [ΔF0 = −12 st, ΔVTL = +3.8 st] (see Fig. 1) under ACE (gray boxes) and SCE (white boxes). The details of the boxplots are as described in Figure 5. ACE indicates advanced combination encoder; JND, just-noticeable-difference; SCE, spectral contrast enhancement; VTL, vocal tract length.

These data indicate that the SCE-induced benefit in SoS intelligibility observed in experiment 1 was not related to the benefit in JNDs (data from experiment 3). In other words, the benefit in SoS intelligibility obtained by participants under the Same Talker and VTL masker conditions did not necessarily stem from SCE improving the participants' sensitivity to the underlying voice differences (F0 and VTL cues) between target and masker speakers.

Taken together, the results from both experiments 1 and 3 reveal that the improvements observed in SoS intelligibility from SCE processing might be due to SCE improving the overall TMR of the signal rather than improving the perception of individual voice cues. This contradicts the postulated hypothesis in this study that the improvement in SoS intelligibility scores from SCE processing arises from SCE improving the sensitivity to the underlying voice cues in the speech signal.

DISCUSSION

The present study was designed to address whether SCE processing would improve SoS intelligibility and comprehension when the competing voices differ from each other parametrically and whether this benefit would arise from SCE improving the sensitivity to the underlying voice cue differences. In line with what has been reported in the literature for speech intelligibility in the presence of SSN (e.g., Baer et al. 1993; Bhattacharya & Zeng 2007; Bhattacharya et al. 2011; Nogueira et al. 2016; Chen et al. 2018) or speech babble maskers (Chen et al. 2018), SCE processing was also found to provide a small yet significant improvement in speech intelligibility in the presence of some single-talker maskers (same female talker and VTL conditions), but not others (F0 and F0+VTL conditions). For masker voice condition F0, the benefit in SoS intelligibility from SCE processing was found to be influenced by large individual variability (see Fig. 6), such that almost half the participants obtained a benefit from SCE relative to ACE, while the other half showed either no difference or a decrement in performance. For masker voice condition F0+VTL, the benefit from SCE appeared to be mitigated by the benefit from the voice difference itself between target and masker. This is because participants seemed to already have gained a benefit in SoS intelligibility scores for this condition relative to when the masker was the same female speaker as the target, thus any potential advantages in intelligibility from SCE were already masked by such a voice benefit.

For SoS comprehension, the data revealed a different picture: no effect of SCE on SoS comprehension accuracy, speed, or drift rate could be observed. There are a number of possibilities as to why this may be the case. One possibility could be that the original female speaker in each corpus was not the same, hence, even though the F0 and VTL differences between the target and masker were always the same in both experiments, the absolute F0 and VTL values were not. However, this explanation seems unlikely because in a similar paradigm

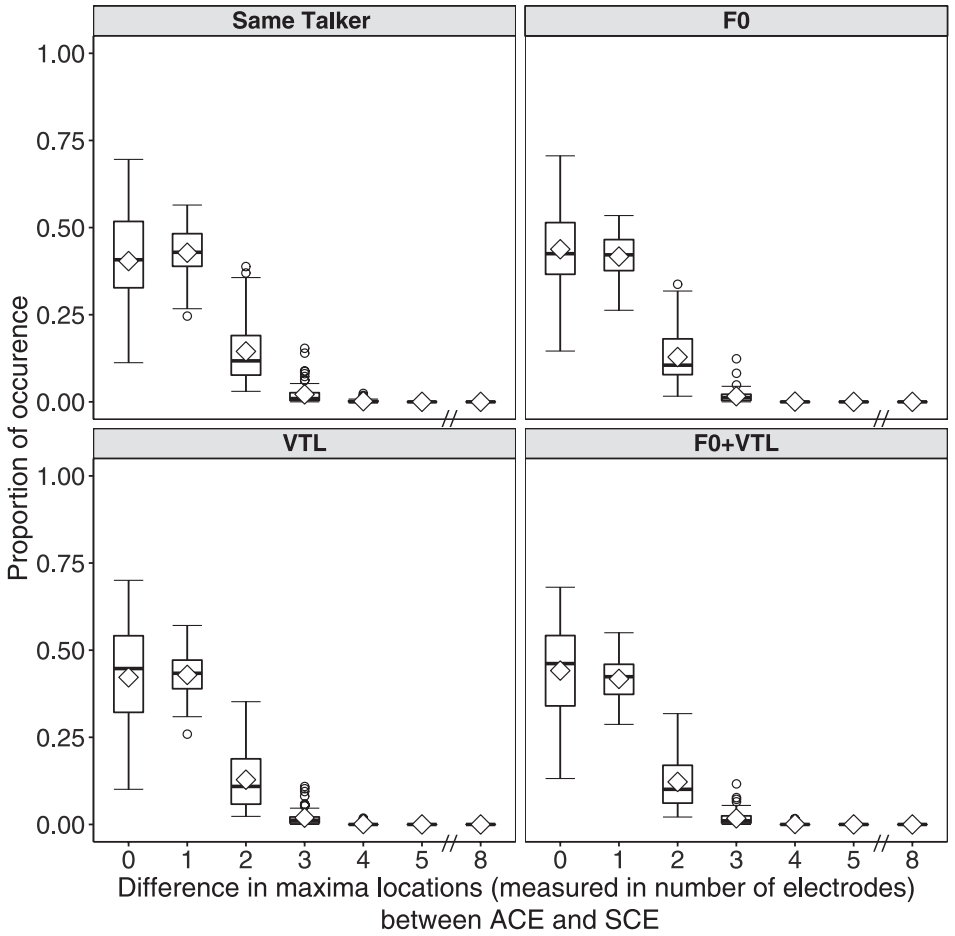

Fig. 9. Proportion of occurrence of differences in maxima (peak) locations between ACE and SCE observed overall temporal frames across all stimuli used in the JND task for all voice conditions. The differences are measured in number of electrodes (peak locations) that differ between ACE and SCE for a given stimulation frame. 0: No difference in peak selection between SCE and ACE; 1: only one peak location is different between the two strategies; 2: Two peak locations are different between the two strategies, and so on. The broken x axis indicates that no difference in peak selection between the two strategies exceeds three electrodes (0 proportion of occurrence until the maximum number of 8 possible peaks is reached). ACE indicates advanced combination encoder; JND, just-noticeable-difference; SCE, spectral contrast enhancement.

(El Boghdady et al. 2019) in which the SoS intelligibility and comprehension material were spoken by different female speakers, the authors still demonstrated that both tasks yielded comparable results for CI users. Another possibility for the differences observed between the SoS intelligibility and comprehension task could be that the sentences in the comprehension task were short and had no additional context. Hence, if a participant misses the subject of the sentence or the resolving word, they were not able to acquire sufficient information to label the sentence as true or false. Moreover, because the intelligibility and comprehension tasks were designed to measure different aspects of speech perception, it may be that SCE influences each of them differently. In that sense, even though SCE may improve the intelligibility of individual words in SoS scenarios, it may be the case that for such sentences without context, as is the case with the SoS comprehension material, this improvement in SoS intelligibility might not be sufficient for overall sentence comprehension.

One key finding of this study was that, in contrast with the previous literature (Stickney et al. 2004; Cullington & Zeng 2008; Pyschny et al. 2011; El Boghdady et al. 2019), CI users in the present study did demonstrate a relatively small, but systematic voice difference benefit in SoS intelligibility. This may be attributed to the nature of the voice manipulations and their direction. For example, in the studies by Pyschny et al. (2011), and El Boghdady et al., (2019), F0 and VTL manipulations were in the direction of shortening VTL and increasing F0, which is the complement of the voice space used in the present study. However, in the studies from Stickney et al. (2004) and Cullington and Zeng (2008), which were performed with real female and male speakers as maskers (not manipulated using software like STRAIGHT), no reported voice benefit was observed. Because the effect of voice space on SoS was not systematically assessed in the present study with the same sample of CI users, it remains an open question whether the voice space influences release from masking in CI users.

It is also not known what the effect of SCE on SoS perception might be for voice differences approaching childlike voices (top-right quadrant of Fig. 1). El Boghdady et al. (2019), in line with the results for Cullington and Zeng (2008), have shown that childlike voice manipulations of the masker appear to yield an additional masking effect for CI users, so the effect of SCE might be more prominent in that voice space. For example, the results of the present study revealed that masker condition

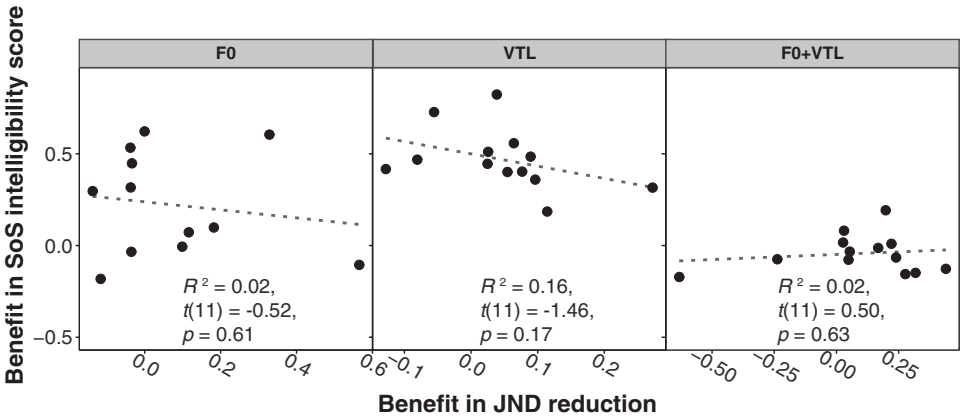

Fig. 10. Correlations between the benefit in SoS intelligibility and voice cue JND reduction from SCE processing. Left panel: Benefit in SoS intelligibility from SCE for masker condition F0 versus benefit as a reduction of F0 JNDs. Middle panel: Benefit in SoS intelligibility from SCE for masker condition VTL versus benefit as a reduction of VTL JNDs. Right panel: Benefit in SoS intelligibility from SCE for masker condition F0+VTL versus benefit as a reduction of F0+VTL JNDs. Positive values denote a benefit from SCE processing, while negative values denote a deficit. JND indicates just-noticeable-difference; SCE, spectral contrast enhancement; SoS, speech-onspeech; VTL, vocal tract length.

F0+VTL yielded an increase in SoS intelligibility scores, which may have masked any further improvements that may have been observed from SCE processing. However, in the previous study of El Boghdady et al. (2019), the authors showed that a similar combination of masker voice in the child space [ΔF0 = +12 st, ΔVTL = −4 st] yielded worse SoS intelligibility for CI users compared to the same talker condition. This reduction in intelligibility could thus be expected to provide space for improvement from SCE processing to manifest. Hence, it may be beneficial to investigate whether SCE could improve such a situation which is initially detrimental to CI users.

The small yet significant benefit in SoS intelligibility from SCE processing in the present study did not appear to stem from improvements in the sensitivity to underlying F0 and VTL cues, as is demonstrated in Figure 10, but rather from improvements in TMR. The data from this study, along with previous findings in the literature, seem to point to the presence of a complex relationship between the TMR, the SCE factor, and the choice of masker voices, and that the particular combination of such parameters as used in this study may have affected the results reported here. In this study, a fixed input TMR of +10 dB was used for the SoS intelligibility task. The literature has provided evidence that the TMR may influence the benefit from SCE. For example, Baer et al. (1993) have shown that for the SVT they used, the effect of spectral enhancement on RTs was larger for moderate TMRs compared to higher TMRs in listeners with hearing loss. On the contrary, for the specific SCE implementation used in the present study, Nogueira et al. (2016) provided simulation data demonstrating that the higher the TMR, the more the background noise root-mean-square intensity was reduced relative to that of the target speech signal for the same SCE factor. The TMR has also been shown to affect the degree of benefit in SoS perception from voice differences between target and masker in a number of studies (e.g., Darwin et al. 2003; Stickney et al. 2004). For very high or very low TMRs, the benefit in SoS perception from voice cue differences becomes minimal, while for intermediate values of the TMR, this benefit increases. These findings support the idea that, in the present study, the benefit in SoS intelligibility obtained from SCE compared with ACE could be dependent on the TMR.

In addition, the SCE factor itself is expected to influence the benefit in TMR relative to ACE, as demonstrated by the simulation results in Figure 4 and those of Nogueira et al., (2016). As the effect of various SCE factors was not assessed in the psychophysics experiments, it may be that SCE factors larger than 0.5 could enhance more important features of speech and attenuate less important ones. However, Baer et al. (1993) suggested that extreme contrast enhancement may negatively impact speech-in-noise perception, so it still remains an open question of whether higher SCE factors would lead to an additional benefit in SoS perception. In addition, because of the large variability in performance observed across participants, it may be worthwhile to investigate customizing the SCE factor for each participant separately.

A final note pertains to the effect of long-term exposure to SCE processing. In the present study, only acute testing in the lab was carried out with the SCE algorithm, and no systematic assessment of long-term exposure or acclimatization was investigated. It is possible that with more prolonged exposure, SCE might yield an additional benefit, as the literature has provided some evidence that, for listeners with hearing loss, the benefit from spectral enhancement could increase with acclimatization (Chen et al. 2018).

The discussion earlier seems to indicate the presence of a complex relationship between the TMR, the SCE factor, the masker voice manipulations, and acclimatization which might influence the benefit in SoS perception from SCE processing. Therefore, a more systematic assessment of this interplay between the aforementioned parameters might be warranted in subsequent studies.

CONCLUSION

This study demonstrated that SCE processing could improve CI users' speech intelligibility in the presence of the same competing talker as the target, or a talker with a lower VTL. This improvement did not appear to arise from SCE enhancing the CI users' sensitivity to F0 and VTL differences between the target and masker speakers per se but rather appeared to arise from inherent improvements introduced to the overall target-to-masker ratio. These findings indicate that SCE could potentially provide some benefit in speech intelligibility for CI users in crowded or

noisy settings. To improve CI users' voice cue perception, other methods should be investigated, such as optimizing the frequency-to-electrode allocation mapping or stimulation techniques to better enhance the representation of spectral cues in the implant.

ACKNOWLEDGEMENTS

The study presented here was jointly funded by Advanced Bionics (AB), the University Medical Center Groningen (UMCG), and the PPP-subsidy of the Top Consortia for Knowledge and Innovation of the Ministry of Economic Affairs, and the German Research Foundation Cluster of Excellence EXC 2177/1 "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 study 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; and the CI participants who took part in this study.

The authors have no conflicts of interest to disclose.

Address for correspondence: Nawal El Boghdady, Department of Otorhinolaryngology, University Medical Center Groningen, Postbus 30.001, 9700 RB Groningen, the Netherlands. E-mail: n.el.boghdady@umcg.nl

OPEN PRACTICES

This manuscript qualifies for an Open Data Badge. The data have been made publically available at https://hdl.handle.net/10411/GPSSP1. More information about the Open Practices Badges can be found at https://journals.lww.com/ear-hearing/pages/default.aspx.

Received February 18, 2019; accepted July 6, 2020.

REFERENCES

Adank, P., Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *J Acoust Soc Am*, *126*, 2649–2659.

Assmann, P., Summerfield, Q. (2004). The perception of speech under adverse conditions. In S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay (Eds), *Speech Processing in the Auditory System*. (pp. 231–308). Springer.

Baddeley, A., Gardner, J. M., Grantham-McGregor, S. (1995). Cross-cultural cognition: Developing tests for developing countries. *Appl Cogn Psychol*, *9*, S173–S195.

Baer, T., Moore, B. C. J., Gatehouse, S. (1993). Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. *J Rehabil Res Dev*, *30*, 49–72.

Başkent, D., Gaudrain, E. (2016). Musician advantage for speech-on-speech perception. *J Acoust Soc Am*, *139*, EL51–EL56.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J Stat Softw*, *67*, 1–48.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B (Methodol)*, *57*, 289–300.

Best, V., Keidser, G., Buchholz, J. M., Freeston K. (2016). Development and preliminary evaluation of a new test of ongoing speech comprehension. *Int J Audiol*, *55*, 45–52.

Bhattacharya, A., Vandali, A., Zeng, F.-G. (2011). Combined spectral and temporal enhancement to improve cochlear-implant speech perception. *J Acoust Soc Am*, *130*, 2951–2960.

Bhattacharya, A., Zeng, F.-G. (2007). Companding to improve cochlear-implant speech recognition in speech-shaped noise. *J Acoust Soc Am*, *122*, 1079–1089.

Bonthuis, M., van Stralen, K. J., Verrina, E., Edefonti, A., Molchanova, E. A., Hokken-Koelega, A. C., Schaefer, F., Jager, K. J. (2012). Use of national and international growth charts for studying height in European children: development of up-to-date European height-for-age charts. *PloS One*, *7*, e42506.

Bosman, A. J., & Smoorenburg, G. F. (1995). Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment. *Audiology, 34*, 260–284.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am*, *109*, 1101–1109.

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. *J Acoust Soc Am*, *120*, 4007–4018.

Cabrera, L., Tsao, F.-M., Gnansia, D., Bertoncini, J., Lorenzi, C. (2014). The role of spectro-temporal fine structure cues in lexical-tone discrimination for French and Mandarin listeners. *J Acoust Soc Am*, *136*, 877–882.

Carlyon, R. P., Shackleton, T. M. (1994). Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? *J Acoust Soc Am*, *95*, 3541–3554.

Chen, J., Moore, B. C., Baer, T., Wu, X. (2018). Individually tailored spectral-change enhancement for the hearing impaired. *J Acoust Soc Am*, *143*, 1128–1137.

Chiba, T., Kajiyama, M. (1941). *The Vowel: Its Nature and Structure*, Tokyo-Kaiseikan.

Cullington, H. E., Zeng, F.-G. (2008). Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects a. *J Acoust Soc Am*, *123*, 450–461.

Darwin, C. J., Brungart, D. S., Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J Acoust Soc Am*, *114*, 2913–2922.

Duquesnoy, A. (1983). Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons. *J Acoust Soc Am*, *74*, 739–743.

El Boghdady, N., Başkent, D., Gaudrain, E. (2018). Effect of frequency mismatch and band partitioning on vocal tract length perception in vocoder simulations of cochlear implant processing. *J Acoust Soc Am*, *143*, 3505–3519.

El Boghdady, N., Gaudrain, E., Başkent, D. (2019). Does good perception of vocal characteristics relate to better speech-on-speech perception in cochlear implant users? *J Acoust Soc Am*, *145*, 417–439.

Fant, G. (1960). Acoustic theory of speech perception. *Mouton, The Hague*.

Festen, J. M., Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J Acoust Soc Am*, *88*, 1725–1736.

Fitch, W. T., Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *J Acoust Soc Am*, *106*, 1511–1522.

Friesen, L. M., Shannon, R. V., Başkent, D., Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *J Acoust Soc Am*, *110*, 1150–1163.

Fu, Q.-J., Nogaki, G. (2005). Noise Susceptibility of Cochlear Implant Users: The Role of Spectral Resolution and Smearing. *J Assoc Res Oto*, *6*, 19–27.

Fu, Q.-J., Shannon, R. V., Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *J Acoust Soc Am*, *104*, 3586–3596.

Fuller, C. D., Gaudrain, E., Clarke, J. N., et al. (2014). Gender Categorization Is Abnormal in Cochlear Implant Users. *J Assoc Res Oto*, *15*, 1037–1048.

Gatehouse, S., Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification. *Br J Audiol*, *24*, 63–68.

Gaudrain, E., Başkent, D. (2015). Factors limiting vocal-tract length discrimination in cochlear implant simulations. *J Acoust Soc Am*, *137*, 1298–1308.

Gaudrain, E., Başkent, D. (2018). Discrimination of voice pitch and vocal-tract length in cochlear implant users. *Ear Hear*, *39*, 226–237.

Goorevich, M., Batty, M. (2005). A new real-time research platform for the Nucleus® 24 and Nucleus® Freedom™ cochlear implants. In *Conference on Implantable Auditory Prostheses (CIAP)*.

Green, D., Swets, J. (1966). *Signal Ddetection Theory and Psychophysics*, Wiley.

Gustafsson, H. A., Arlinger, S. D. (1994). Masking of speech by amplitude-modulated noise. *J Acoust Soc Am*, *95*, 518–529.

Hahlbrock, D. K.-H. (1953). Über Sprachaudiometrie und neue Wörterteste. *Archiv f. Ohren-, Nasen- u. Kehlkopfheilkunde*, *162*, 394–431.

Hochmair-Desoyer, I., Schulz, E., Moser, L., Schmidt, M. (1997). The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users. *Am J Otol*, *18*, S83.

Ives, D. T., Smith, D. R. R., Patterson, R. D. (2005). Discrimination of speaker size from syllable phrases. *J Acoust Soc Am*, *118*, 3816–3822.

Kawahara, H., Irino, T. (2005). Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In P. Divenyi (Ed.) *Speech Separation by Humans and Machines*. (pp. 167–180). Springer.

Kiessling, J., Pichora-Fuller, M. K., Gatehouse, S., Stephens, D., Arlinger, S., Chisolm, T., Davis, A. C., Erber, N. P., Hickson, L., Holmes, A., Rosenhall, U., von Wedel, H. (2003). Candidature for and delivery of audiological services: special needs of older people. *Int J Audiol*, *42*, 92–101.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J Acoust Soc Am*, *49*, 467–477.

Licklider, J. (1954). "Periodicity" pitch and "place" pitch. *J Acoust Soc Am*, *26*, 945–945.

Lieberman, P., Blumstein, S. E. (1988). Source-filter theory of speech production. In *Speech Physiology, Speech Perception, and Acoustic Phonetics*. (pp. 34–50). Cambridge University Press.

Lo, S., Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Front Psychol*, *6*:1171.

Loizou, P. C., Poroy, O. (2001). Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners. *J Acoust Soc Am*, *110*, 1619–1627.

Meister, H., Fürsen, K., Streicher, B., Lang-Roth, R., Walger, M. (2016). The use of voice cues for speaker gender recognition in cochlear implant recipients. *J Speech Lang Hear Res*, *59*, 546–556.

Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *J Assoc Res Oto*, *9*, 399–406.

Müller, J. (1848). *The Physiology of the Senses, Voice, and Muscular Motion, with the Mental Faculties*..., Taylor, Walton & Maberly.

Nelson, P. B., Jin, S.-H. (2004). Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners. *J Acoust Soc Am*, *115*, 2286–2294.

Nelson, P. B., Jin, S.-H., Carney, A. E., Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *J Acoust Soc Am*, *113*, 961.

Nogueira, W., Büchner, A., Lenarz, T., Edler, B. (2005). A psychoacoustic NofM-type speech coding strategy for cochlear implants. *EURASIP J Appl Signal Process*, *2005*, 3044–3059.

Nogueira, W., Rode, T., Buechner, A. (2016). Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants. *J Acoust Soc Am*, *139*, 728–739.

Oxenham, A. J. (2008). Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants. *Trends Amplif*, *12*, 316–331.

Oxenham, A. J., Simonson, A. M., Turicchia, L., et al. (2007). Evaluation of companding-based spectral enhancement using simulated cochlear-implant processing. *J Acoust Soc Am*, *121*, 1709–1716.

Pachella, R. G. (1974). The interpretation of reaction time in human information processing research. In B.H. Kantowitz (Ed.), *Human Information Processing: Tutorials in Performance and Cognition*. Erlbaum Associates.

Pals, C., Sarampalis, A., Beynon, A., Stainsby, T., Başkent, D. (2020). Effect of spectral channels on speech recognition, comprehension, and listening effort in cochlear-implant users. *Trends Hear*, 24, 2331216520904617. https://doi.org/10.1177/2331216520904617

Pals, C., Sarampalis, A., van Rijn, H., Başkent, D. (2015). Validation of a simple response-time measure of listening effort. *J Acoust Soc Am*, *138*, EL187–EL192.

Peterson, G. E., Barney, H. L. (1952). Control methods used in a study of the vowels. *J Acoust Soc Am*, *24*, 175–184.

Pisoni, D. B., Manous, L. M., Dedina, M. J. (1987). Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. *Comput Speech Lang, 2*, 303–320.

Pyschny, V., Landwehr, M., Hahn, M., Walger, M., von Wedel, H., Meister, H. (2011). Bimodal hearing and speech perception with a competing talker. *J Speech Lang Hear Res*, *54*, 1400–1415.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rana, B., Buchholz, J. M., Morgan, C., Sharma, M., Weller, T., Appaiah Konganda, S., Shirai, K., Kawano, A. (2017). Bilateral versus unilateral cochlear implantation in adult listeners: Speech-on-speech masking and multitalker localization. *Trends Hear*, *21*, 1–15.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol Rev*, *85*, 59–108.

Ratcliff, R., Smith, P. L., Brown, S. D., McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends Cogn Sci*, *20*, 260–281.

Schaffrath Rosario, A., Schienkiewitz, A., Neuhauser, H. (2011). German height references for children aged 0 to under 18 years compared to WHO and CDC growth charts. *Ann Hum Bio*, *38*, 121–130.

Schouten, J., Bekker, J. (1967). Reaction time and accuracy. *Acta Psychol*, 27, 143–153.

Smith, D. R. R., Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *J Acoust Soc Am*, *118*, 3177–3186.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., Irino, T. (2005). The processing and perception of size information in speech sounds. *J Acoust Soc Am*, *117*, 305–318.

Stevens, K. N., House, A. S. (1955). Development of a quantitative description of vowel articulation. *J Acoust Soc Am*, *27*, 484–493.

Stickney, G. S., Zeng, F.-G., Litovsky, R., Assmann, P. (2004). Cochlear implant speech recognition with speech maskers. *J Acoust Soc Am*, *116*, 1081–1091.

Turicchia, L., Sarpeshkar, R. (2005). A bio-inspired companding strategy for spectral enhancement. *IEEE Trans Speech Audio Process*, *13*, 243–253.

Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., Henry, B. A. (2004). Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing. *J Acoust Soc Am*, *115*, 1729–1735.

Wagenmakers, E.-J., Van Der Maas, H. L., Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychon Bull & Rev*, *14*, 3–22.

Wang, S., Xu, L., Mannell, R. (2011). Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners. *J Assoc Res Oto*, *12*, 783–794.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol*, *41*, 67–85.

Zaltz, Y., Goldsworthy, R. L., Kishon-Rabin, L., Eisenberg, L. S. (2018). Voice discrimination by adults with cochlear implants: the benefits of early implantation for vocal-tract length perception. *JARO*, *19*, 193–209.