Research Article

Effect of Dual-Carrier Processing on the Intelligibility of Concurrent Vocoded Sentences

Frédéric Apoux,^a Brittney L. Carter,^b and Eric W. Healy^b

Purpose: The goal of this study was to examine the role of carrier cues in sound source segregation and the possibility to enhance the intelligibility of 2 sentences presented simultaneously. Dual-carrier (DC) processing (Apoux, Youngdahl, Yoho, & Healy, 2015) was used to introduce synthetic carrier cues in vocoded speech.

Method: Listeners with normal hearing heard sentences processed either with a DC or with a traditional single-carrier (SC) vocoder. One group was asked to repeat both sentences in a sentence pair (Experiment 1). The other group was asked to repeat only 1 sentence of the pair and was provided additional segregation cues involving onset asynchrony (Experiment 2).

Results: Both experiments showed that not only is the “target” sentence more intelligible in DC compared with SC,

but the “background” sentence intelligibility is equally enhanced. The participants did not benefit from the additional segregation cues.

Conclusions: The data showed a clear benefit of using a distinct carrier to convey each sentence (i.e., DC processing). Accordingly, the poor speech intelligibility in noise typically observed with SC-vocoded speech may be partly attributed to the envelope of independent sound sources sharing the same carrier. Moreover, this work suggests that noise reduction may not be the only viable option to improve speech intelligibility in noise for users of cochlear implants. Alternative approaches aimed at enhancing sound source segregation such as DC processing may help to improve speech intelligibility while preserving and enhancing the background.

Vocoder processing is a technique used to represent acoustic signals as the sum of a limited number of amplitude-modulated sinewaves or noise bands, with the modulations being derived from the corresponding frequency bands in the original signal. Because vocoded sounds transmit primarily the amplitude envelope fluctuations of the signal in a limited number of channels, many studies have used this processing as a tool to study the contribution of amplitude fluctuations in speech recognition. These studies have shown that substantial intelligibility of speech in quiet is possible, provided that at least four to eight spectral channels are used (e.g., Apoux & Bacon, 2008; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). In other words, a severe reduction in spectral information

and a complete elimination of temporal fine structure (TFS) cues do not severely impair speech intelligibility in quiet. Accordingly, it is believed that envelope cues convey sufficient information to support accurate speech recognition.

However, vocoded speech becomes difficult to understand when a competing signal is simultaneously present, especially if this competing signal is amplitude modulated, such as a competing talker. The significant intelligibility deterioration has led to the assumption that some cues critical for understanding speech in noise are not conveyed by the envelope. More specifically, it has been suggested that temporal envelopes alone do not provide the cues necessary to extract a target signal from a mixture of sounds. Considering that spectral and TFS cues are severely distorted or entirely discarded in vocoded processing, it is reasonable to assume that at least one of these cues is potentially important for sound source segregation.

The role of spectral cues in speech recognition in noise has been indirectly illustrated in many studies using diverse approaches. One approach involved vocoder processing and consisted of manipulating the subjects' spectral resolution by varying the number of channels. These studies

^aDepartment of Otolaryngology, The Ohio State University Wexner Medical Center, Columbus

^bDepartment of Speech and Hearing Science, The Ohio State University, Columbus

Correspondence to Frédéric Apoux: fred.apoux@gmail.com

Editor-in-Chief: Frederick (Erick) Gallun

Editor: Jennifer Lentz

Received June 16, 2017

Revision received December 5, 2017

Accepted June 11, 2018

https://doi.org/10.1044/2018_JSLHR-H-17-0234

Disclosure: The authors have declared that no competing interests existed at the time of publication.

generally showed a large effect of the number of spectral channels in noise (Dorman, Loizou, Fitzke, & Tu, 1998; Friesen, Shannon, Baskent, & Wang, 2001; Fu, Shannon, & Wang, 1998; Qin & Oxenham, 2003). This contrasts with the relatively limited effect observed in quiet. Another approach also consisted of manipulating the subjects' spectral resolution or by evoking excitation patterns in the normal ear that resemble those in the impaired ear. Consistent with the results of vocoder studies, these studies showed that smearing the spectra of speech is not largely detrimental in quiet, at least for smearing that simulates auditory filters up to six times broader than normal (e.g., Baer & Moore, 1993). The intelligibility of speech in noise, however, is affected even when small amounts of spectral smearing are introduced (e.g., Baer & Moore, 1994). Both vocoder processing and smearing approaches demonstrate the consistent result that spectral cues are more critical for speech recognition in noise than in quiet.

The apparent relationship between speech recognition in noise and spectral resolution has often been considered evidence of a role of spectral cues in streaming. However, recent developments in our understanding of the mechanisms underlying speech recognition in noise offer alternative interpretations of this relationship. In particular, some models of speech recognition in noise such as the glimpsing model (e.g., Apoux & Healy, 2013; Cooke, 2006) suggest that the increased intelligibility in noise typically observed with increasing frequency selectivity is not related to a better representation of spectral cues. Instead, the influence of fine spectral resolution is primarily associated with the ability to ignore or reject the spectrotemporal regions most affected by noise (e.g., Apoux, Yoho, Youngdahl, & Healy, 2013). It is assumed that the auditory system continuously divides the acoustic environment into small time–frequency units for subsequent processing. Because of the fluctuating properties of most natural sounds, the signal-to-noise ratio (SNR) in these units often differs from the average SNR estimated from the broadband sound mixture over longer periods. This difference tends to increase as the size of the time–frequency units decreases, therefore improving the ability to reject noise (i.e., to reject units with an unfavorable SNR). Because the spectral width of these units is essentially determined by the width of the auditory filters, it becomes apparent how poor frequency resolution may affect a listener's ability to reject noise. Accordingly, the glimpsing model assumes that the influence of spectral resolution is primarily a noise rejection effect.

The role of TFS cues in extracting a target from the background is less clear. To our knowledge, there is only limited and indirect evidence of such a role. First, the results of vocoder studies can be considered as supporting the contribution of TFS cues to sound source segregation. Indeed, the intelligibility of vocoded speech in noise remains below that of unprocessed speech even with as many as 20 channels (e.g., Friesen et al., 2001), suggesting that poor frequency selectivity is not the only factor limiting recognition in noise. Recently, Apoux and Healy (2013) evaluated

more directly the role of TFS cues in sound source segregation. This study assessed the relative importance of the target and background TFSs. The authors manipulated the number and nature of TFSs used to convey the envelopes of two competing sentences. These envelopes were mixed and imposed on TFSs extracted from the isolated target speech, on TFSs extracted from the isolated background speech, or on TFSs extracted from the mixed target-plus-background. The results of this study showed that preserving only the target TFS is equivalent to preserving only the background TFS, indicating that the nature of the TFS has little influence on intelligibility (and segregation). More importantly, it showed that preserving the mixed TFSs is by far the most advantageous condition, suggesting that a critical factor for effective speech recognition in noise is the unique association between an envelope and a TFS. These findings were interpreted according to the glimpsing model, and the authors concluded that TFS cues are primarily used to isolate and group together the time–frequency units containing pertinent information about the target signal (i.e., the glimpses).

This conclusion was subsequently evaluated in Apoux, Youngdahl, Yoho, and Healy (2015). The authors proposed a technique to transmit synthetic TFS information in vocoded speech (i.e., independent arbitrary carriers), based on previous studies showing that the TFS is not the primary carrier of speech information and therefore its preservation is not necessary (Apoux & Healy, 2013; Apoux et al., 2013). This approach is in clear contrast to current attempts that manipulate the carrier to somehow reflect the TFS of the incoming sounds to preserve the original TFS information (e.g., Arnoldner et al., 2007; Riss, Arnoldner, Baumgartner, Kaider, & Hamzavi, 2008). An apparent advantage of using arbitrary carriers is that the original TFS, which may be difficult to derive and encode, is not needed.

In the technique proposed by Apoux et al. (2015), TFS segregation cues were restricted to a difference in carrier rates. Consistent with Apoux et al. (2015) and the description of acoustic signals in the time domain (see Equations 1 and 5 in Apoux, Millman, Viemeister, Brown, & Bacon, 2011), carrier cues are considered as TFS-based cues in this study. Apoux et al. (2015) reasoned that the use of different rates may signal which time–frequency units are glimpses and which are noise with enough salience to significantly improve speech intelligibility in noise. Accordingly, two independent and arbitrary carriers, one conveying the target speech envelope and one conveying the background envelope, were used in the vocoded sound mixture. These carriers consisted of two pulse trains that differed only in their fixed rate. Apoux et al. (2015) observed a systematic intelligibility benefit of “dual-carrier” (DC) processing over traditional single-carrier (SC) vocoding, ranging from 24% to 50% points. This benefit, obtained with carriers that were not derived from the TFS of the encoded sounds, strongly supports the role of carrier cues, and more generally, TFS-based cues, in sound source segregation.

A central question not addressed in the previous study is that of target selection or background intelligibility. This

is a central aspect of auditory scene analysis as it allows for a differentiation between enhancement of true sound source segregation and enhancement of target intelligibility. For instance, target intelligibility may be substantially improved by suppressing the background. Obviously, this approach cannot be considered a form of sound source segregation enhancement. Although unlikely, the findings of Apoux et al. (2015) may have been caused by unseen factors that predominantly increased the intelligibility of the target rather than enhanced sound source segregation. Therefore, to truly establish the facilitatory effect of DC processing and the general role of TFS-based cues in auditory scene analysis, it is important to demonstrate an improvement in intelligibility for either signal in the sound mixture. In other words, DC processing also needs to enhance access to the background.

Accordingly, this study was designed to evaluate the effect of DC processing on the intelligibility of target and background speech signals. Two complementary approaches were employed. In Experiment 1, recognition scores were collected for two sentences presented concurrently. Repetition was allowed so that participants could process both sentences in the same stimulus. In Experiment 2, participants were cued to repeat either the target or the “background” and repetition was not allowed. On the basis of Apoux et al. (2015), it was hypothesized that target and background intelligibility would be similarly enhanced.

Experiment 1: Intelligibility of the Target and Background in the Same Sentence Pair

Methods

Design

The goal of Experiment 1 was to directly assess the intelligibility of target and background speech signals by asking listeners with normal hearing (NH) to report both sentences in a pair. The DC carrier-rate combination was chosen so that similar intelligibility would be achieved with either rate (see Apoux et al., 2015). Because it is difficult, if not impossible, to process two simultaneous sentences even in ideal nonprocessed conditions, participants were allowed to repeat each stimulus up to 15 times. Two SC conditions, one for each carrier rate, were also included.

Participants

This study was performed on human participants with approval from The Ohio State University Institutional Review Board. Eight participants with NH participated in Experiment 1. NH was defined as having pure-tone air-conduction thresholds of 20 dB or better, at octave frequencies from 250 to 8000 Hz (American National Standards Institute, 2004, 2010). The participants were aged 19–27 years (six women and two men; mean age = 21.5 years). They received course credit for their participation.

Stimuli and Processing

Sixty-four sentences from the IEEE corpus (IEEE, 1969), each containing five scoring keywords and all spoken

by the same male talker, were used. The sampling rate for all the stimuli was 22050 Hz (including Experiment 2). Sentences were chosen to have a similar duration (2800–2924 ms) to limit asynchrony or fringe cues (Bacon & Grantham, 1992; Darwin, 1981, 1984; Oxenham & Dau, 2001; Rasch, 1978), although up to a 124-ms fringe may have been present. Sentences were summed into pairs so that the fringe would always occur at the end of the stimulus, further limiting its influence (e.g., Zera & Green, 1993). Although they are referred to here as “target” and “background” sentences, this determination was made arbitrarily and so the distinction is simply for clarity.

Three processing conditions were compared: one DC condition with carrier rates of 150 and 250 pulses per second (pps) and two SC conditions with carrier rates of 150 or 250 pps. The pulse duration was set to 10 samples. Because nulls in the overall shape of the spectrum occur at intervals $1/t$, where t is the duration of the pulse, the intelligibility of our stimuli may have been reduced. However, all SC and DC conditions should have been equally affected, making all the comparisons appropriate. Pulse train carriers were obtained by concatenating the same pulse as many times as needed to match the exact duration of the stimuli. Each pulse train started with a silence corresponding to half the interpulse gap. Therefore, the pulses were synchronous when using the same rates but not when using different rates, and synchronicity was preserved across channels. The interpulse duration depended on the carrier rate. DC processing was implemented as the sum of two independent SC vocoders, one employing a pulse train carrier at 150 pps and the other at 250 pps. It should be noted at this point that the two concurrent sentences need to be separated from each other before DC processing. Although suitable solutions exist to extract a target from the background (e.g., noise reduction), the purpose of this study was to evaluate the intelligibility of both target and background speech signals after DC processing. Therefore, it was necessary to exclude extraneous factors such as the influence of the extraction process. Accordingly, a priori knowledge of the target and background sentences (i.e., “ideal separation”) was used.

Each SC vocoder was implemented as follows: Stimuli were filtered into 10 contiguous frequency bands ranging from 80 to 7563 Hz. Filtering (third-order Butterworth bandpass) was cascaded and performed in both the forward and reverse directions so that the process produced zero phase distortion, yielding a total effective filter order of 24 (Apoux & Healy, 2009). Each band was three normal equivalent rectangular bandwidth (Glasberg and Moore, 1990). Therefore, the bandwidths from the lowest band to the highest band were as follows: 118, 162, 225, 310, 427, 592, 816, 1127, 1556, and 2150 Hz. This approach roughly simulates the number of independent channels typically available in listeners with cochlear implants (CIs; e.g., Croghan, Duran, & Smith, 2017; Friesen et al., 2001), potentially allowing additional implications. The envelope was extracted from each band by half-wave rectification and low-pass filtering at c/f_m (single-pass eighth-order Butterworth, 48-dB/octave

roll-off; see Healy & Steinbach, 2007), with *c/fm* corresponding to half the bandwidth (in Hertz) of the normal equivalent rectangular bandwidth centered on that band (Apoux & Bacon, 2008). The filtered envelopes were then used to modulate fixed-rate pulse train carriers having rates of 150 or 250 pps. The modulated pulse trains were bandpass filtered using the same procedures used to create the spectral bands, to restrict their frequency range to the bandwidth of the corresponding channel, and then summed over all channels to produce the SC stimulus.

Filtering can affect the nature of a pulse train, especially if the bandwidth of the filter is relatively narrow. These unavoidable effects represent a necessary trade-off when assessing multiple-carrier processing in subjects with NH. The consequences of the current pulse-train filtering can be summarized as follows. The lowest channel (80–198 Hz) did not always contain carrier energy, meaning that a signal, whether target or background, was not always present. The carriers were more tonal than pulsatile in Channels 2 and 3 (198–360 and 360–585 Hz), potentially allowing some form of spectral segregation. This tonal characteristic faded away with increasing center frequency (due to increasing bandwidth), with Channel 4 (585–895 Hz) resembling a pulse train more than a tone. In the six remaining channels (895 Hz and above), the carriers generally preserved their pulselike characteristics. The main consequence is that the implications for CI users should be interpreted with caution as the same effects would not necessarily be observed in electrical stimulation. It should be noted, however, that previous work involving CI subjects showed that these listeners can benefit by as much as 40% points from a stimulation approach involving multiple carriers (e.g., Mc Laughlin, Reilly, & Zeng, 2013). Therefore, the present results should remain generally representative of what might be possible in CI users.

In the SC condition, the target- and background-sentence envelopes were each imposed on separate but identical, in-phase, pulse trains. After summation, the target and background envelopes shared a single carrier. In the DC conditions, the resulting stimulus was a sound mixture made up of two modulated pulse trains. To ensure similar intelligibility for the target and background, the signal-to-noise ratio (SNR) for all conditions was set to 0 dB.

It should be noted that the current SC implementation may differ slightly from a traditional vocoder in which the complex envelope is used to modulate a single pulse train. We chose to use two parallel SC vocoders because this implementation is more comparable with DC processing. However, an informal pilot study suggested that the intelligibility of the sum of the envelopes is comparable with that of the envelope of the sum.

Procedure

The 64 sentences were randomly arranged into 32 pairs, half of which were presented in the DC condition and the other half were presented in the SC conditions. For the SC conditions, sentences were again divided into two sets of eight sentence pairs, with each set corresponding to one

carrier rate. Sentences were blocked by processing condition (DC or SC) and then by carrier rate in SC. The SC conditions were heard first by half of the participants, and the DC condition was heard first by the other half. For the SC conditions, the carrier rate that was tested first alternated between participants. The order of the sentence pairs was not randomized, so that sentence pair-to-condition correspondence was balanced. Participants were tested individually in a double-walled, sound-attenuated booth with the experimenter present. Test stimuli were presented diotically through Sennheiser HD 280 Pro circumaural headphones using computers equipped with Echo Gina 3G digital-to-analog converters. The level of each sentence pair was set to play at 65 dBA at each earphone using a flat plate coupler (Larson Davis AEC 101) and Type 1 sound-level meter (Larson Davis 824).

A short practice was provided first, which was composed of 20 IEEE sentences not used for formal testing, presented as single vocoded sentences with no background. Half of the sentences were processed at the 150-pps carrier rate, and the other half were processed at the 250-pps carrier rate. During formal testing, participants were instructed that they would hear two sentences played simultaneously and were to repeat back as much of each sentence as possible, guessing when unsure. Spoken responses were scored for keywords correct by the experimenter. Participants pressed the keyboard spacebar to play the sentence pair and were permitted to play each sentence pair until all 10 keywords (five per sentence) were correctly repeated or up to a maximum of 15 presentations. They were then allowed to move on to the next stimulus pair.

Results

Individual-subject intelligibility scores are shown in Figure 1. It should be noted that scores in the SC conditions correspond to the percentage of correct keywords for both sentences, whereas scores in the DC conditions correspond to the percentage of correct keywords for each sentence separately. As can be seen, sentence intelligibility in DC exceeded that in SC for all participants and for both carrier rates, confirming the advantage of using multiple carriers to transmit a mixture of sounds. This is despite the fact that subjects were allowed up to 15 repetitions, therefore providing ample opportunities to process each sentence in a pair. Subjects used 12–15 repetitions ($M = 14.39$) in SC and 4–15 repetitions ($M = 13.74$) in DC. The individual improvement from SC to DC ranged from 34% to 61% points for the 250-pps carrier rate and 38%–58% points for the 150-pps carrier rate. More importantly, target and background sentences were equally intelligible in DC, suggesting that it is possible to improve speech intelligibility in noise for at least two speech signals within a mixture, without having to suppress one of those signals or even determine which is the signal of interest.

Percent correct scores were transformed into rationalized arcsine units before statistical analysis (Studebaker, 1985). A two-way repeated-measures analysis of variance

Figure 1. Individual-subject sentence recognition scores in single-carrier (SC) and dual-carrier (DC) processing. Target carrier rates in pulses per second (pps) are indicated in parentheses.

was performed using the factors processing and carrier rate. Only the main effect of processing was significant, $F(1, 7) = 980.43, p < .001$, revealing that intelligibility in DC ($M = 70.1\%$, $SD = 8.0\%$) was significantly higher than that in SC ($M = 21.8\%$, $SD = 7.5\%$). The main effect of rate, $F(1, 7) = 0.62, p > .05$, and the interaction of processing and rate, $F(1, 7) = 0.13, p > .05$, were nonsignificant. This verifies the initial postulates that neither carrier rate would be more advantageous in DC (69% at 150 pps vs. 71% at 250 pps) and that neither carrier rate would be more advantageous in SC (21% at 150 pps vs. 23% at 250 pps).

Experiment 2: Relative Intelligibility of the Target and Background: Fringe Effects

Methods

Design

In Experiment 1, direct evidence was provided that DC processing allows listeners to effectively process both target and background speech signals in a vocoded sound mixture. To examine if it was a facilitation of sound segregation resulting from carrier cues provided by the dual carriers, the participants were required to process both sentences in the same pair. Experiment 2 was designed to provide a complementary view of target and background intelligibility in DC conditions. In this experiment, participants were asked to repeat only one of two sentences in a pair, and stimulus repetition was not permitted.

Thus far, we have evaluated the benefit from DC processing using synchronous, gated sentences. This approach eliminates most of the asynchrony cues that listeners with NH use and that may also be used when listening to SC vocoded speech. A possible consequence of eliminating

asynchrony cues is to exaggerate the contribution of carrier cues to sound source segregation as they may become the only cues available. Accordingly, Experiment 2 was designed to preserve asynchrony cues to some extent. One sentence in each pair was temporally delayed relative to the other, resulting in a forward-sentence fringe. On the basis of previous work regarding onset asynchrony, sentence recognition should be facilitated by temporally delaying one sentence relative to the other. It remains unclear, however, if this facilitating effect would reduce the benefit of DC processing and therefore diminish the role of carrier cues.

Participants

Ten participants with NH who were not included in Experiment 1 participated in Experiment 2. NH was defined as in Experiment 1. They were aged 19–24 years (nine women and one man; mean age = 20.3 years) and received course credit for their participation.

Stimuli and Processing

Target sentences were again selected from the IEEE corpus, and all were produced by the same male talker. The word “say,” taken from a sentence produced by the same male talker, was appended to the beginning of each target sentence. This cue word duration was approximately 500 ms, and it was always processed to match the processing of the target sentence. Sentences from the AzBio corpus (Spahr et al., 2012) were used as background. Each background sentence was produced by a male or female talker, randomly selected. To limit confusion, background sentences were time reversed. Three fringe conditions were tested and are illustrated in Figure 2. In one fringe condition, cue + target and background sentences were gated on and off simultaneously (no fringe, NF). In another fringe condition, the cue + target sentence preceded the background sentence by the 500-ms cue word (target fringe, TF). As a result, only the five scored keywords of the target were presented during the background sentence. In the last fringe

Figure 2. Schematic of the three fringe arrangements of Experiment 2.

condition, the background sentence preceded the cue + target sentence by 500 ms (background fringe, BF). To ensure that the background sentence had the exact desired duration for each target sentence and condition, the time-reversed background sentences were concatenated and a random start point was selected for each trial.

Six processing conditions were compared: one unprocessed (UNP) condition, one SC condition, and four DC conditions. In the UNP condition, the target and background sentences were summed and presented to participants without further processing. This condition represents the “gold standard” performance of listeners with NH. In the remaining five conditions, target and background sentences were vocoded independently before summation using two independent SC vocoders, as described in the Methods section of Experiment 1.

In DC, the target and background envelopes were imposed on pulse trains of different rates, resulting in each envelope being conveyed by its own carrier. Two carrier rate combinations were used. One combination employed rates of 250 pps for the target and 150 pps for the background. As previously mentioned and confirmed in Experiment 1, this balanced (BAL) combination results in equal target and background intelligibility. The second combination employed carrier rates of 250 pps for the target and 300 pps for the background and was chosen to yield maximum target intelligibility (MAX), based on the results of Apoux et al. (2015). These two combinations were also tested while inverting the sentence-to-carrier-rate correspondence. In other words, the third and fourth combinations employed carrier rates of 150 pps for the target and 250 pps for the background (BAL⁻¹) or 300 pps for the target and 250 pps for the background (MAX⁻¹), respectively. The SC condition was created using a carrier rate of 250 Hz for both the target and the background. As in Experiment 1, target and background were matched in level for an SNR of 0 dB.

Procedure

Participants were tested individually with the experimenter present in a double-walled, sound-attenuated booth. Stimuli were presented diotically through Sennheiser HD 280 Pro circumaural headphones. Testing was performed using custom MATLAB scripts running on computers equipped with Echo Gina 3G digital-to-analog converters.

A short practice was provided that included 20 IEEE sentences not used during formal testing. Participants first heard four UNP sentences in quiet. The remaining 16 sentences were DC processed using the MAX combination, to provide the participants the modest practice on what was likely the “easiest” condition. The DC-processed sentences were presented first in quiet, and then the SNR was set to 16, 8, and 0 dB after every four sentences.

Formal testing consisted of six blocks of 12 sentence pairs each, with each block corresponding to one processing condition. Blocks were presented in a random order for each participant. Before each of the five vocoded conditions,

participants listened to six sentences processed using an SC vocoder having a carrier rate matching that of the target. This was done to ensure that the participants attended to the target sentence rather than the background sentence. In addition, the first two sentences of each block of 12 served as practice and were therefore discarded. Participants were instructed that they would hear two sentences played simultaneously and were to repeat back as much of the cued target sentence as possible while ignoring the background sentence and to guess when unsure. No repetition was permitted. The participants’ spoken responses were scored for keywords correct in the target sentence by the experimenter. The presentation level for all stimuli was set to 65 dBA, as in Experiment 1.

Results

Figure 3 shows mean percent correct performance as a function of the processing. The parameter is the fringe arrangement. As expected, performance was highest in the UNP condition, averaging 73%. Consistent with the results of Apoux et al. (2015) and Experiment 1, DC processing yielded a systematic benefit over SC processing. This benefit varied across processing conditions from 6% points in MAX⁻¹ (BF) to 34% points in MAX (TF), which is consistent with Apoux et al. (2015). Also consistent with Apoux et al. (2015) is the result that sentence intelligibility was systematically lower in MAX⁻¹ than in any other DC condition and that DC was always better than SC. A few discrepancies were noted. For instance, although BAL and BAL⁻¹ were expected to yield similar performance, participants achieved slightly better scores in BAL. In addition, whereas the best performance was expected to be achieved

Figure 3. Mean sentence recognition scores as a function of the target/background rate in pulses per second (pps): unprocessed (UNP), single-carrier (250/250), and dual-carrier (250/300 [MAX], 300/250 [MAX⁻¹], 250/150 [BAL], 150/250 [BAL⁻¹]). The parameter is fringe arrangement: no fringe (NF), target fringe (TF), and background fringe (BF). Error bars indicate 1 SE.

in the MAX conditions, the two highest DC scores were obtained in BAL (NF and BF). Despite these slight discrepancies, the present data are generally consistent with the predictions from Apoux et al. (2015) and therefore can be considered representative of the effect of DC processing.

From Figure 3, it is apparent that intelligibility was not substantially affected by the addition of a fringe. In UNP, adding a fringe to either the target or the background slightly affected performance, reducing intelligibility by 3%–4% points. Although the fringe effect was also limited in the vocoded conditions, a slightly different pattern was observed. Generally, the TF yielded a drop in performance of 1%–10% points (average = 7% points). In contrast, a BF generally improved intelligibility by 2%–4% points, with the exception of the BAL condition in which a drop of 3% points was observed. Overall, the addition of a forward fringe on either signal had a limited influence on performance and produced marginally different patterns between UNP and vocoded conditions.

A two-way repeated-measures analysis of variance was performed including the factors of processing (UNP, SC, MAX, MAX⁻¹, BAL, and BAL⁻¹) and fringe (NF, TF, and BF). Before this analysis, percent correct scores were transformed to rationalized arcsine units, as for Experiment 1. As expected, the analysis yielded a significant main effect of processing, $F(5, 45) = 87.96, p < .001$. Post hoc comparisons (Holm–Sidak) indicated that all the processing conditions differed significantly from each other ($p < .03$), except for BAL and MAX ($p = .51$). More surprisingly, the main effect of fringe was also found to be significant, $F(2, 18) = 31.10, p < .001$. Post hoc comparisons (Holm–Sidak) indicated that NF and BF did not differ significantly from each other ($p = .27$), but both differed from TF ($p < .001$). The interaction of processing and fringe was nonsignificant, $F(10, 90) = 1.57, p > .05$.

Discussion

One of the primary advantages of a multicarrier approach such as DC processing is the restoration of some of the TFS-based cues abolished by vocoder processing. Our previous work showed that using independent carriers could improve the intelligibility of the target signal while preserving the background. On the basis of this work and the relatively subjective distinction between target and background signals in our DC approach, it was hypothesized that speech intelligibility would be similar for both target and background signals. In Experiment 1, the ability to process either sentence in a simultaneous pair was directly examined. Results from Experiment 1 confirmed that the intelligibility of the target can be substantially improved when isolated on its own independent carrier and, more importantly, that access to the background is similarly enhanced, provided that the influence of the carrier rate is taken into account. In other words, there is no significant difference between target and background intelligibility in DC processing when an appropriate carrier rate

combination is used.¹ The current results support the hypothesis that TFS-based cues can support segregation by indicating which units within the overall signal correspond to the target (which units are glimpses) and which units correspond to the background. This support is derived from the fact that the speech information presented in SC and DC is the same, and the difference involves the pairing of each individual envelope with a different arbitrary carrier, rather than their combination on the same arbitrary carrier.

Because it is difficult to process two simultaneous sentences, participants were allowed to hear each sentence pair multiple times in Experiment 1. Although this procedure does not reflect real-world situations, it does provide the unique opportunity to unambiguously evaluate the relative intelligibility of each signal present in the sound mixture. It is likely that allowing multiple repetitions of each sentence pair increased performance. However, the fact that participants had the opportunity to repeat each sentence pair in all conditions helps provide a fair comparison between SC and DC. In addition, it is interesting to note that repetition itself is not sufficient to alleviate the difficulties associated with speech recognition in noise. Indeed, although participants may have benefited to some extent from this repetition in SC, segregation remained very challenging overall and performance stagnated around 20%.

Experiment 2 also provided some indication of the intelligibility of the background, as each rate was separately tested as the target carrier. Overall, Experiment 2 confirmed that the access to both sentences in a pair can be substantially enhanced by DC processing. It also showed that both sentences can be equally intelligible when selecting the appropriate rates.

Surprisingly, the fringe effect reported in previous studies was not observed here. We hypothesized that participants would initially focus on the sentence starting first (i.e., the fringe). They would then be required to either maintain their focus on that first sentence or switch to the temporally delayed sentence. We also hypothesized that the TF would significantly improve intelligibility by facilitating the target carrier identification.² However, as the results show, this was never the case.

First, performance did not differ significantly between BF and NF. A possible explanation is that the BF did facilitate segregation. However, the 500-ms cue word may have acted as a BF in NF, resulting in facilitation comparable with BF. Indeed, one may reasonably assume that these 500 ms provided the participants with ample time to focus their attention on the target sentence, therefore nullifying the fringe advantage.

¹The target and background signals are most likely equally intelligible in SC processing as well, albeit to a much lesser extent.

²In particular, the presence or absence of the cue word in the fringe might be a cue to determine if the leading sentence was the target or not.

Second, performance dropped significantly when the fringe was added to the target (compared with NF and BF), seemingly contradicting our above interpretation. The contradiction, however, is only in appearance. Quite possibly, there was a beneficial influence of the TF in that it immediately revealed which carrier conveyed the target sentence. As a consequence, the participants should have been able to focus their attention on the target sentence well before the background sentence even appeared. Unfortunately, this benefit may have been offset by an involuntary shift in attention occurring at the onset of the background sentence. Indeed, as explained by Escera and Malmierca (2014), “Not only largely salient, novel, ‘unique’ stimuli are capable of driving attention involuntarily, but also small stimulus changes occurring occasionally in an otherwise repetitive sound sequence, that is, ‘deviant’, or contextually novel stimuli, are capable of attracting attention” (p. 111). Studies involving the mismatch negativity paradigm provide perhaps the best illustration of such a tendency to detect and shift attention to unexpected novel stimuli. For example, Schröger (1996) demonstrated that a mismatch-negativity-eliciting tone, which deviated in frequency from the standard stimuli in the unattended channel of a selective-attention task, reduced behavioral performance for a subsequent target occurring in the attended channel. Interestingly, this effect was restricted to a short latency (200 ms) between the unattended deviant and the attended target and disappeared when the latency was extended to 560 ms. In BF, participants appropriately switched focus to the target as it was the novel signal, and therefore intelligibility was not affected. In TF, the novel signal happened to be the background. The appearance of the background possibly drew the participants’ attention away from the target, at least for a brief moment. Because this appearance often coincided by design with the presentation of the first keyword, the participants may have missed a small part of this first keyword, resulting in a small detrimental effect of the TF.

Consistent with our previous work, this study demonstrates that DC processing yields a systematic improvement over traditional SC vocoder processing. Although substantial, the improvement observed here (up to 61% points) was not sufficient to match the level of performance achieved with UNP. A similar gap between DC and UNP was observed by Apoux et al. (2015). Part of this gap may be attributed to the considerable practice participants have on unprocessed speech. Another interpretation is that the carrier cues conveyed in DC conditions are not sufficient to fully compensate for distortions introduced by vocoder processing. For instance, the nature of the carriers used in DC may contribute to the remaining gap in performance between this condition and UNP. Indeed, the use of different fixed carrier rates, as implemented here, does not capture the distinctive frequency modulations (FMs) of the original carriers. Although the importance of these FMs may be limited in the presence of a nonharmonic masker (Apoux et al., 2015), it has been shown that removing these FMs may adversely affect intelligibility in noise (Brown, Helms

Tillery, Apoux, Doyle, & Bacon, 2015; Carroll, Tiaden, & Zeng, 2011). Therefore, it may be postulated that at least part of the performance gap between DC and UNP could be attributed to the lack of FM cues.

Perhaps, a more significant factor is the reduction in frequency selectivity associated with DC processing, as the number of frequency bands was set to 10 to mimic the general number of available channels demonstrated by listeners with CIs. As a consequence, frequency selectivity differed greatly between DC and UNP as participants benefited from normal selectivity in the latter. According to some glimpsing models (e.g., Apoux & Healy, 2013), the auditory system reconstructs a representation of a target speech signal by combining information from regions “dominated” by this target speech. The role of normal frequency selectivity is critical in this process because it allows the decomposition of the mixture of sounds into a large number of independent frequency regions. As stated in the Introduction, the narrower the regions, the more their SNR will deviate from the overall SNR. In other words, the higher the frequency selectivity, the better the auditory system will be at isolating even the smallest regions with usable speech information. Because frequency selectivity was much higher in UNP than in DC, it should not be surprising that performance was also higher in the former condition. Overall, we believe that practice and, more importantly, adding more channels in DC processing could potentially close the DC–UNP gap.

In addition to increased clarity regarding the role of TFS-based cues in speech recognition in noise, the current study illustrates another benefit of DC processing—that it preserves the richness of the acoustic environment. Not only can listeners better understand the target signal, but they can also have intelligible access to the background. The similarity between vocoder and CI processing has led us to consider the potential application of DC processing to CIs. Although our results do not provide evidence that CI users could make use of multiple carriers,³ they strongly suggest that the SC approach may be one of the primary factors responsible for the poor speech intelligibility in noise commonly displayed by these users. Accordingly, it may be of interest to discuss some potential benefits of an approach inspired by the DC technique for CIs.

Streaming is a particularly valuable listening tool in multi-talker environments where efficient communication requires that listeners first identify which of multiple speech signals is coming from the talker of interest, then attend to that talker of interest by ignoring background speech signals, and then switch attention between talkers when needed. The two primary factors contributing to CI recipients’ streaming difficulty are most likely reduced frequency selectivity and the loss of TFS information

³However, based on previous work investigating temporal pitch perception in CI recipients (Kong, Deeks, Axon, & Carlyon, 2009), it may be assumed that using two low pulse rates (< 300 pps) or combining one low rate with one high rate (e.g., 250 and 900 pps) may elicit two distinct pitch percepts.

associated with CI processing. Given the relationship between frequency selectivity and glimpsing, it is easy to understand how having 10 or fewer independent frequency channels, such as recipients of CIs, can be a severe hindrance. For recipients of CIs, this process is further limited by the complete loss of TFS cues. As suggested by Apoux and Healy (2013) and Apoux et al. (2013) and confirmed here, TFS may be critical for identifying which portions of a sound mixture provide glimpses of a target signal. Without TFS information, individuals with CIs must rely on weaker and more inefficient segregation cues such as envelope periodicity (Hong & Turner, 2009).

A potential benefit of an implementation of DC in CIs is evidently to improve target intelligibility when background sounds are present. However, because DC processing requires that the target signal first be isolated from the background, one may reasonably question the value of preserving the background, rather than simply discarding it. It is important to note that the goal of DC processing is to provide the listener with a mixture of sound sources processed in a way that allows the auditory system to naturally extract the desired signal from the mixture, just as listeners with normal hearing are able to do with unprocessed speech in noise. This logic is different from that of traditional noise reduction. Listeners who only have access to vocoded-like stimuli, such as CI users, may aspire to be more aware of their surroundings provided that intelligibility of the speech is satisfactory. Measurements⁴ in our laboratory formally assessed the “cost of preserving the background,” which refers to the difference in intelligibility at a given SNR between background suppressed (traditional noise reduction) and background preserved and transmitted on an independent carrier (DC). Results showed there is no cost of preserving the background for SNR improvements that may be achieved by currently implemented noise reduction systems (e.g., Brons, Houben, & Dreschler, 2015). Accordingly, we concluded that there should be no cost of preserving the background in DC processing when using existing noise reduction technology.

It is important to note that the goal of DC processing is not to match the performance of listeners with NH in quiet. The goal is to match the performance of listeners with NH in noise and to preserve the richness of the acoustic environment. Accordingly, although preserving the background can reduce intelligibility in some situations, this cost should not be considered a drawback of DC processing per se. Rather, speech intelligibility in noise is simply lower than that in quiet. The present data further illustrate this advantage by showing that not only is it possible to significantly enhance target speech intelligibility while preserving the background, but also it is possible to make this background intelligible and accessible to the listener. When the background contains another

talker, this speech can be just as intelligible as the target. When two or more simultaneous talkers share the background carrier, they may not all be highly understandable, but at worst, the DC strategy will still result in improved target intelligibility while providing an awareness of the background.

Another advantage of DC processing illustrated here is to limit the consequences of erroneous target selection, an obligatory step in noise reduction. When using noise reduction in two-talker environments, a prediction must be made as to which signal is of interest to the user and which signal is noise. Although this step is also present in DC processing to separate the target and the background, the consequences of selecting the “wrong” target are understandably not as critical because both signals are presented (and intelligible if composed of two voices). In contrast to suppressing the background, the DC approach offers the advantage of placing the user at the center of the decision-making process in that she or he ultimately decides what signal is of interest. This is only possible because, as demonstrated here, the background can be as intelligible as the target.

Acknowledgments

Execution of this study was supported in part by R01 DC08594, and article preparation was supported in part by R01 DC15521, both to author E. W. H.

References

- American National Standards Institute.** (2004). *Methods for manual pure-tone threshold audiometry* (ANSI S3.21-2004 [R2009]). New York, NY: Acoustical Society of America.
- American National Standards Institute.** (2010). *Specification for audiometers* (ANSI S3.6-2010). New York, NY: Acoustical Society of America.
- Apoux, F., & Bacon, S. P.** (2008). Differential contribution of envelope fluctuations across frequency to consonant identification in quiet. *The Journal of the Acoustical Society of America*, 123, 2792–2800.
- Apoux, F., & Healy, E. W.** (2009). On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence. *Hearing Research*, 255, 99–108.
- Apoux, F., & Healy, E. W.** (2013). A glimpsing account of the role of temporal fine structure information in speech recognition. In B. C. J. Moore, R. D. Patterson, I. M. Winters, R. P. Carlyon, & H. E. Gockel (Eds.), *Basic aspects of hearing: Physiology and perception* (pp. 119–126). New York, NY: Springer.
- Apoux, F., Millman, R. E., Viemeister, N. F., Brown, C. A., & Bacon, S. P.** (2011). On the mechanisms involved in the recovery of envelope information from temporal fine structure. *The Journal of the Acoustical Society of America*, 130, 273–282.
- Apoux, F., Yoho, S. E., Youngdahl, C. L., & Healy, E. W.** (2013). Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners. *The Journal of the Acoustical Society of America*, 134, 2205–2212.

⁴Portions of these measurements were presented in Apoux and Healy (2015), “Dual-carrier vocoder processing: Cost of preserving the background,” at the 42nd Annual American Auditory Society Scientific and Technology Meeting, Podium Paper V.H., Scottsdale, USA.

- Apoux, F., Youngdahl, C. L., Yoho, S. E., & Healy, E. W. (2015). Dual-carrier processing to convey temporal fine structure cues: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, 138, 1469–1480.
- Arnoldner, C., Riss, D., Brunner, M., Durisin, M., Baumgartner, W. D., & Hamzavi, J. S. (2007). Speech and music perception with the new fine structure speech coding strategy: Preliminary results. *Acta Otolaryngologica*, 127, 1298–1303.
- Bacon, S. P., & Grantham, D. W. (1992). Fringe effects in modulation masking. *The Journal of the Acoustical Society of America*, 91, 3451–3455.
- Baer, T., & Moore, B. C. J. (1993). Effects of spectral smearing on the intelligibility of sentences in noise. *The Journal of the Acoustical Society of America*, 94, 1229–1241.
- Baer, T., & Moore, B. C. J. (1994). Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. *The Journal of the Acoustical Society of America*, 95, 2277–2280.
- Brons, I., Houben, R., & Dreschler, W. A. (2015). Acoustical and perceptual comparison of noise reduction and compression in hearing aids. *Journal of Speech, Language, and Hearing Research*, 58, 1363–1376.
- Brown, C. A., Helms Tillery, K., Apoux, F., Doyle, N. M., & Bacon, S. P. (2015). Shifting fundamental frequency in simulated electric-acoustic listening: Effects of F0 variation. *Ear and Hearing*, 37, 18–25.
- Carroll, J., Tiaden, S., & Zeng, F.-G. (2011). Fundamental frequency is critical to speech perception in noise in combined acoustic and electric hearing. *The Journal of the Acoustical Society of America*, 130, 2054–2062.
- Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119, 1562–1573.
- Croghan, N. B. H., Duran, S. I., & Smith, Z. M. (2017). Re-examining the relationship between number of cochlear implant channels and maximal speech intelligibility. *The Journal of the Acoustical Society of America*, 142, EL537–EL543.
- Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology*, 33, 185–207.
- Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *The Journal of the Acoustical Society of America*, 76, 1636–1647.
- Dorman, M. F., Loizou, P. C., Fitzke, J., & Tu, Z. (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. *The Journal of the Acoustical Society of America*, 104, 3583–3585.
- Escera, C., & Malmierca, M. S. (2014). The auditory novelty system: An attempt to integrate human and animal research. *Psychophysiology*, 51, 111–123.
- Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America*, 110, 1150–1163.
- Fu, Q. J., Shannon, R. V., & Wang, X. S. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *The Journal of the Acoustical Society of America*, 104, 3586–3596.
- Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47, 103–138.
- Healy, E. W., & Steinbach, H. M. (2007). The effect of smoothing filter slope and spectral frequency on temporal speech information. *The Journal of the Acoustical Society of America*, 121, 1177–1181.
- Hong, R. S., & Turner, C. W. (2009). Sequential stream segregation using temporal periodicity cues in cochlear implant recipients. *The Journal of the Acoustical Society of America*, 126, 291–299.
- IEEE. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17, 225–246.
- Kong, Y.-Y., Deeks, J. M., Axon, P. R., & Carlyon, R. P. (2009). Limits of temporal pitch in cochlear implants. *The Journal of the Acoustical Society of America*, 125, 1649–1657.
- Mc Laughlin, M., Reilly, R. B., & Zeng, F. G. (2013). Rate and onset cues can improve cochlear implant synthetic vowel recognition in noise. *The Journal of the Acoustical Society of America*, 133, 1546–1560.
- Oxenham, A. J., & Dau, T. (2001). Modulation detection interference: Effects of concurrent and sequential streaming. *The Journal of the Acoustical Society of America*, 110, 402–408.
- Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *The Journal of the Acoustical Society of America*, 114, 446–454.
- Rasch, R. A. (1978). Perception of simultaneous notes such as in polyphonic music. *Acta Acustica United with Acustica*, 40, 21–33.
- Riss, D., Arnoldner, C., Baumgartner, W. D., Kaider, A., & Hamzavi, J. S. (2008). A new fine structure speech coding strategy: Speech perception at a reduced number of channels. *Otology & Neurotology*, 29, 784–788.
- Schröger, E. (1996). Neural mechanism for involuntary attention shifts to changes in auditory stimulation. *Journal of Cognitive Neuroscience*, 8, 527–539.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303–304.
- Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., ... Cook, S. (2012). Development and validation of the AzBio sentence lists. *Ear and Hearing*, 33, 112–117.
- Studebaker, G. A. (1985). A rationalized arcsine transform. *Journal of Speech and Hearing Research*, 28, 455–462.
- Zera, J., & Green, D. M. (1993). Detecting temporal onset and offset asynchrony in multicomponent complexes. *The Journal of the Acoustical Society of America*, 93, 1038–1052.