# Using Zebra-speech to study sequential and simultaneous speech segregation in a cochlear-implant simulation

Etienne Gaudrain[a] and Robert P. Carlyon
*MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, CB2 7EF Cambridge, United Kingdom*

(Received 14 December 2011; revised 10 November 2012; accepted 14 November 2012)

Previous studies have suggested that cochlear implant users may have particular difficulties exploiting opportunities to glimpse clear segments of a target speech signal in the presence of a fluctuating masker. Although it has been proposed that this difficulty is associated with a deficit in linking the glimpsed segments across time, the details of this mechanism are yet to be explained. The present study introduces a method called Zebra-speech developed to investigate the relative contribution of simultaneous and sequential segregation mechanisms in concurrent speech perception, using a noise-band vocoder to simulate cochlear implants. One experiment showed that the saliency of the difference between the target and the masker is a key factor for Zebra-speech perception, as it is for sequential segregation. Furthermore, forward masking played little or no role, confirming that intelligibility was not limited by energetic masking but by across-time linkage abilities. In another experiment, a binaural cue was used to distinguish the target and the masker. It showed that the relative contribution of simultaneous and sequential segregation depended on the spectral resolution, with listeners relying more on sequential segregation when the spectral resolution was reduced. The potential of Zebra-speech as a segregation enhancement strategy for cochlear implants is discussed.
 [http://dx.doi.org/10.1121/1.4770243]

PACS number(s): 43.71.Ky, 43.66.Dc, 43.66.Mk, 43.66.Pn [LD] Pages: 502–518

## I. INTRODUCTION

Cochlear implants (CIs) have proved successful at restoring speech understanding in quiet but noisy environments remain extremely challenging. Speech perception in the presence of competing talkers is a complex phenomenon but it can be decomposed into a combination of the following simpler mechanisms. Considering a short segment of the mixture, the target signal must be segregated from the interfering sound within this segment, which is generally referred to as (a) *simultaneous segregation* (Assmann and Summerfield, 1990; Bregman, 1990; de Cheveigné, 1999; Vestergaard *et al*., 2009). The successive segments must also be linked across time to form a continuous percept. This across-time linkage can be referred to as (a) *sequential segregation* or *streaming* (Bregman, 1990; Dorman *et al*., 1975; Gaudrain *et al*., 2007, 2008; van Noorden, 1975; Nooteboom *et al*., 1978). As speech presents some natural interruptions (stops and pauses), segments of the target signal can be remote in time. This is aggravated when the target signal is interrupted by an overlapping competing sound. If simultaneous segregation fails in these segments, the listener will "miss" some parts of the target signal, and so the perceived segments will have to be linked across longer gaps than if no segments were missed. Depending on the success of these two segregation mechanisms, the resulting phonetic information that can be extracted might be degraded, and inferring the meaning of the sentence can be non-trivial.

Simultaneous and sequential segregation can be expected to be affected differently by loss of frequency selectivity, as it occurs, for instance, in CIs. When two sound events are simultaneous, and therefore substantially overlap in time, energetic masking will only be avoided if the two sounds do not overlap in the spectral domain. Reduced frequency selectivity, by smearing the spectral content of the competing sounds, increases the amount of spectral overlap, and therefore increases the energetic masking occurring in this situation. The notched-noise method for auditory filter bandwidth estimation, for instance, directly exploits this phenomenon (Glasberg and Moore, 1990; Patterson, 1976). Reducing this masking in the implant would require a restoration of frequency selectivity. On the other side, in the scenario of sequential segregation, the competing sounds do not overlap in time. Reduced frequency selectivity may affect the salience of some segregation cues, which is known to be the crucial factor for streaming (Moore and Gockel, 2002). However, at least in the absence of substantial forward masking, direct energetic masking of consecutive segments should not occur and should not worsen with frequency-selectivity loss. To the extent that this is true, an improvement in sequential segregation could be achieved by increasing the saliency of the segregation cue.

This rather theoretical decomposition finds some experimental support in various correlational studies involving hearing-impaired listeners. Summers and Leek (1998) did not find any significant correlation between simultaneous segregation abilities and concurrent speech intelligibility in normal-hearing (NH) and hearing-impaired listeners, whereas correlations between sequential segregation and concurrent speech perception have been observed in other studies (Gaudrain *et al*., 2012; Mackersie *et al*., 2001). Despite these

[a]Author to whom correspondence should be addressed. Current address: University of Groningen, University Medical Center Groningen, ENT, PO Box 30.001, 9700 RB Groningen, Netherlands. Electronic mail: e.p.c.gaudrain@umcg.nl

observations, the roles of each of these mechanisms in concurrent speech perception were not studied together, and therefore the way they are affected by hearing impairment remains unclear. Recent studies on glimpsing in CI users (Gnansia *et al*., 2010; Nelson and Jin, 2004; Nelson *et al*., 2003; Qin and Oxenham, 2003; all reviewed in Secs. I A–I C), also suggested that both simultaneous segregation and "across-time linkage" were limiting intelligibility for these listeners. Although in these studies both mechanisms were present and the interpretation proposed by the authors established a contrast between them, the two mechanisms were not directly investigated and their contribution could not be compared.

The present study primarily focuses on sequential segregation and aims to investigate the factors driving across-time linkage of speech segments in simulated CIs. The main novelty is that, by introducing a new method called *Zebra-speech* that allows the neutralization of the simultaneous segregation mechanism while preserving intelligibility, we were able to compare the contributions from the two segregation mechanisms. Experiment 1 was used to validate the method and to select the value of its only parameter to be used in the remaining experiments. Experiments 2a and 2b tested the assumption that sound segregation with Zebra-speech is limited only by failures in across-time linkage, and in particular that energetic masking caused by forward masking does not play a major role. Experiments 3a and 3b used Zebra-speech to investigate the role of various binaural cues for the across-time linkage of speech segments. Sections I A, I B, and I C present reviews of the literature on glimpsing and across-time linkage in CI recipients, on simultaneous segregation in CI users, and on sequential segregation in these listeners.

### A. Glimpsing and across-time linkage in (simulated) CI listeners

When two competing signals fluctuate both in time and frequency—as, for example, in speech—listeners can glimpse relatively clear spectro-temporal portions of the target signal where they are not energetically masked (Cooke, 2003, 2006; Howard-Jones and Rosen, 1993; Miller and Licklider, 1950). Identifying which spectro-temporal portions, or "cells," can be glimpsed, and re-assembling them to reconstruct a meaningful signal are necessary for comprehension. Cells pertaining to a given source that occupy different frequency regions but that overlap in time are grouped using simultaneous segregation mechanisms. Cells that do not overlap in time are grouped using sequential segregation mechanisms. Cells that partially overlap in time will require a combination of simultaneous and sequential segregation. A loss in intelligibility can be caused by failures to group segments either across time or frequency. The term "dip-listening" (Buus, 1985) is generally associated with a version of the glimpsing theory where temporal segments (vertical stripes in the spectrogram) are glimpsed instead of spectro-temporal portions (cells in the spectrogram). This concept of dip-listening is therefore directly related to the problems of across-time linkage and sequential segregation, which are the main focus of our study. Note that in the rest of this article, the term "glimpsing" refers to purely temporal glimpsing, i.e., listening in temporal dips, and is therefore directly associated with the sequential segregation mechanism.

Experiments with CI users suggest that they may have particular difficulties exploiting glimpsing or dip-listening strategies (Nelson *et al*., 2003; Qin and Oxenham, 2003; but see also Bernstein and Brungart, 2011). These studies have shown that the ability of CI listeners to benefit from masker fluctuations when identifying target sentences is reduced compared to that of listeners with NH, which could be interpreted as a deficit in sequential grouping. However, to maintain the same overall target-to-masker ratio (TMR), the most intense portions of a fluctuating masker have to be more intense than the equivalent steady-state masker; hence producing more energetic masking and making simultaneous segregation more difficult during the interruptions, which could have a more adverse effect in CIs than in NH. To avoid this complication, perception of interrupted (or interleaved) speech can be compared in CI and NH listeners. Using this method, Nelson and Jin (2004) and Gnansia *et al*. (2010) concluded that aggravated energetic masking did not explain all the deficits of CI listeners, and that the reduced performance might therefore also be caused by increased non-energetic masking. In these two studies, results were attributed to a deficit in across-time tracking and integration of disjoint segments of the target signal. In a recent study, Kwon *et al*. (2012) measured speech reception in noise maskers designed to either promote or suppress simultaneous energetic-masking release. They found that the difference in performance between these two conditions was smaller for CI users than for NH listeners, which is, again, not consistent with a deficit in simultaneous segregation only.

A similar conclusion was reached in a different context where NH listeners had to listen to spatially-separated concurrent sentences. Ihlefeld and Shinn-Cunningham (2008) reduced energetic masking by passing the target and masker signals through sine-wave vocoders with interleaved frequency bands. They then studied the effect of TMR and spatial separation between the target and the masker. A detailed analysis of the pattern of response errors led to the conclusion that the spatial-separation benefit was caused by release from energetic masking at lower TMRs, while at higher TMRs it was due to either improvement in across-time linkage (streaming) or improvement in selective attention.

The effect of listening through a CI has also been studied directly on simultaneous and sequential segregation themselves in studies that are reviewed in Secs. I B and I C.

### B. Simultaneous segregation

Simultaneous segregation in speech has often been studied using the concurrent vowel or concurrent syllable paradigms (de Cheveigné, 1999; Culling and Darwin, 1993; Scheffers, 1983; Vestergaard and Patterson, 2009). In NH listeners, segregation performance is driven by TMR, and by differences in location, fundamental frequency ($F0$), and other vocal characteristics such as vocal-tract length. However, in real or simulated CI users, no $F0$-related benefit has been observed due to the relatively poor representation of

this feature in CIs (Luo and Fu, 2009; Luo *et al*., 2009; Qin and Oxenham, 2005). This leaves only TMR, location, and possibly other cues related to speaker identity, as potential factors that influence performance. Because the stimuli are usually presented monaurally, and because CIs do not transmit fine spectral detail, the limits on simultaneous segregation by CI users can be regarded as largely due to energetic masking (Carlyon *et al*., 2007).

The extent to which simultaneous segregation is involved in longer (e.g., sentence length) competing utterances depends on how the TMR varies over time. When the "instantaneous" TMR is largely positive, simultaneous segregation is trivial. On the contrary, when the TMR is largely negative, simultaneous segregation is likely to be impossible. So simultaneous segregation really only occurs and plays a role when the TMR is within some intermediate range encompassing 0 dB. Figure 1 shows distributions of TMR evaluated in 40 ms chunks of sentences from a British version of the Coordinate Response Measure (CRM) corpus (Bolia *et al*., 2000; Kitterick *et al*., 2010) in various interferers. The interferers are composed of sentences from the IHR Sentence Lists corpus concatenated so as to match the duration of the target sentence and mixed at 0 dB overall TMR. Distributions built from 1000 such mixtures show that when the interferer consists of a single talker, the TMR is greater than 9.5 dB or smaller than −9.3 dB 50% of the time, and is either greater than 18.7 dB or smaller than −20.0 dB 25% of the time. So when the number of competing speakers is small, then even when the nominal TMR is 0 dB, it should be borne in mind that a large proportion of the signal segments show a clear dominance of one speaker or the other. In the case of a single competing talker, as in the case of gated noise interferers, sequential segregation is thus expected to play a significant role. In contrast, when four speakers are masking the target sentence, the median positive TMR is only 4.1 dB, and opportunities to glimpse a clear portion of the target signal are relatively scarce. Note that these considerations are not only relevant for behavioral experiments but also for studies investigating the neural correlates of speech segregation (e.g., Mesgarani and Chang, 2012).

### C. Sequential segregation

Previous studies of sequential segregation have suggested that the salience of the difference between two sources determines how the auditory scene is perceptually organized (Moore and Gockel, 2002). However, most studies on sequential segregation involved artificial steady-state sounds with few high-level features. It is then unclear how this conclusion extends to rich and dynamic auditory objects. Studies using modulated noise interferers might therefore yield different results than those involving a competing talker. This seems to be the case even in (simulated) CIs despite the reduced spectral resolution (Qin and Oxenham, 2003). To avoid this issue, Gnansia *et al*. (2010) filled the gaps of some periodically interrupted sentences with segments from a competing talker, extending a technique used by Miller and Licklider (1950), to the study of CI recipients. They observed that filling the gaps with speech greatly reduced identification performance for NH listeners as well as for real and simulated CIs, indicating that interruption (i.e., the loss of information caused by deleting portions of the target) alone does not account for all the deficits that listeners might encounter when hearing two concurrent speakers.

Addressing more basic aspects of the sequential segregation mechanism, there exists competing evidence on whether CI listeners do or do not have preserved stream segregation abilities (Chatterjee *et al*., 2006; Hong and Turner, 2006, 2009; see also Cooper and Roberts, 2009). Whether

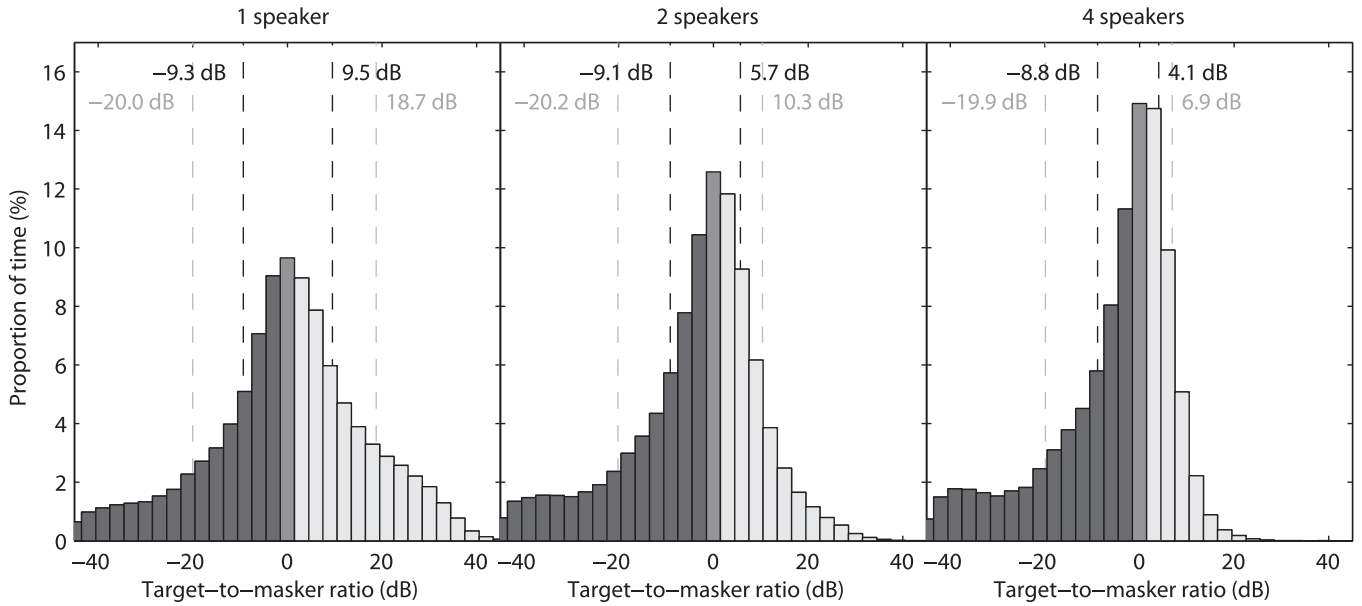


FIG. 1. Distribution of the TMR evaluated on 40 ms chunks of 1000 instances of a target sentence mixed with interferers composed of 1, 2, and 4 talkers, respectively (see text for details). Vertical black dashed lines represent medians for positive and negative TMRs, gray dashed lines represent 25-percentiles.

this is the case or not, the reduced ability to extract strong and salient pitches in CIs will likely reduce the use of streaming in speech-like situations (e.g., Gaudrain *et al.*, 2008). More generally, what does seem clear is that the substantial current spread between electrodes is likely to at least hinder streaming, as has also been shown for simultaneous segregation (Carlyon *et al.*, 2007). However, when speech material is used and if forward masking remains limited, energetic masking should not limit the access to the linguistic information in a sequential configuration.

A potential limitation in all these studies on sequential segregation, as in the studies on interrupted or interleaved speech, is that the various sources involved came on and off according to an unnaturally regular pattern. One exception is the seminal study by Miller and Licklider (1950) where little or no effect of regularity was observed on the perception of interrupted speech. However, it has been observed that regularity has an effect on the streaming of interleaved melodies (Devergie *et al.*, 2010). It cannot be ruled out that, in the case of impaired hearing where little other segregation cues are available, regularity could play a greater role in the perception of interrupted, interleaved, or concurrent speech. The investigation method proposed in the current study produces irregular interruptions to avoid this potential caveat.

In summary, both simultaneous and sequential segregation are important in concurrent speech perception, at least when the number of competing talkers is relatively small, and both mechanisms are impaired in CI hearing. Simultaneous segregation seems limited by spectral resolution. Improving simultaneous segregation would consequently require improved spectral resolution. In contrast, sequential segregation seems mostly related to target-masker discriminability (a hypothesis tested in Experiment 2), which can potentially be enhanced by amplifying existing cues or by introducing new cues. A better understanding of the role of sequential segregation in ecological situations could therefore lead to a direct benefit for CI users, a topic discussed in Sec. VI.

In Sec. II, a new method—Zebra-speech—is introduced along with the general methods that are common to all the experiments presented. In Experiment 1 the only parameter of the Zebra-speech method was varied, and an optimal value was selected and used in all subsequent experiments. The objective of Experiment 2 was to assess whether forward masking affects Zebra-speech intelligibility. Finally, Experiment 3 used Zebra-speech to investigate the role of various binaural cues in the sequential segregation of speech. All experiments made use of noise-band vocoders to simulate the effect of listening through a CI, and in particular to mimic the spectral resolution available to CI users.

## II. ZEBRA-SPEECH AND GENERAL METHOD

Zebra-speech is built from two separate sentences by keeping, at each time, only the most intense of the two signals, as evaluated by the root-mean-square (RMS) in chunks of a fixed duration (see Fig. 2). The name is suggested by the fact that alternate vertical "stripes" of the spectrogram may represent, respectively, a segment of the target and of the masker sentence (cf. the "checkerboard speech" of Howard-Jones and Rosen, 1993). Transitions between the target and the masker were smoothed by convolving the alternation pattern (a binary vector indicating which voice is selected) with a 5 ms Hann window. The effect of chunk duration on intelligibility was investigated in Experiment 1. A property of Zebra-speech is that when the two signals have the same overall RMS, i.e., mixed with a TMR of 0 dB, and the same modulation statistics, the resulting Zebra-speech signal will also present a TMR of 0 dB and the two signals will on average be selected the same proportion of the time.

In addition to Zebra processing, we also mixed the target and masker conventionally by simply summing the two signals. This method is henceforth referred to by the shorthand *Donkey-speech* because, unlike Zebra-speech where the alternating black and white stripes of the animal represent the alternation between the two voices, Donkey-speech is uniformly composed of an equal sum of the two voices just like donkeys are, generally, uniformly gray.

A basic idea underlying the use of Zebra-speech is that by only sacrificing the portions of the signal that are the most challenging to retrieve (having a negative TMR), intelligibility is minimally affected, thus allowing the study of sequential mechanisms in ecological yet controlled conditions. In all of the following experiments, the target and masker sentences were mixed in Zebra or Donkey mode and were then noise-band vocoded to simulate a CI.

### A. Stimuli

#### 1. Target and masker signals

In all the experiments, the target sentences were from a British version of the CRM corpus (Bolia *et al.*, 2000; Kitterick *et al.*, 2010). The sentences from this corpus all have the same structure: "Ready ⟨*call-sign*⟩ go to ⟨*color*⟩ ⟨*number*⟩ now." The number ranges from one to eight, and the colors can be "green," "red," "white," or "blue." Eight different call-signs are available and were all used but were irrelevant for the task given to the participants. The sound files were sampled on 16 bits at 44.1 kHz and equalized in RMS. Only two of the eight speakers available in the corpus were used: A male speaker ($M3$, average $F0 = 102$ Hz) and a female speaker ($F3$, average $F0 = 232$ Hz). On each trial a random sentence was selected and mixed with the masker signal.

The maskers were derived from the IHR Sentence List corpus (MacLeod and Summerfield, 1987), recorded from another male speaker (average $F0 = 139$ Hz). The original sentences were sampled on 16 bits at 22 050 Hz, and were resampled to 44.1 kHz prior to any manipulation. Semantic content was obfuscated by time-reversing the sentences to avoid semantic interference. This made the task easier and removed the need to decide "who said what." On each trial, randomly selected sentences were concatenated to produce a signal longer than the target CRM sentence. A random segment of the same duration as the target sentence was then selected from this concatenated signal. The onset and offset of the segment were then smoothed using 10-ms ramps. Finally, the level of the masker signal was adjusted so that

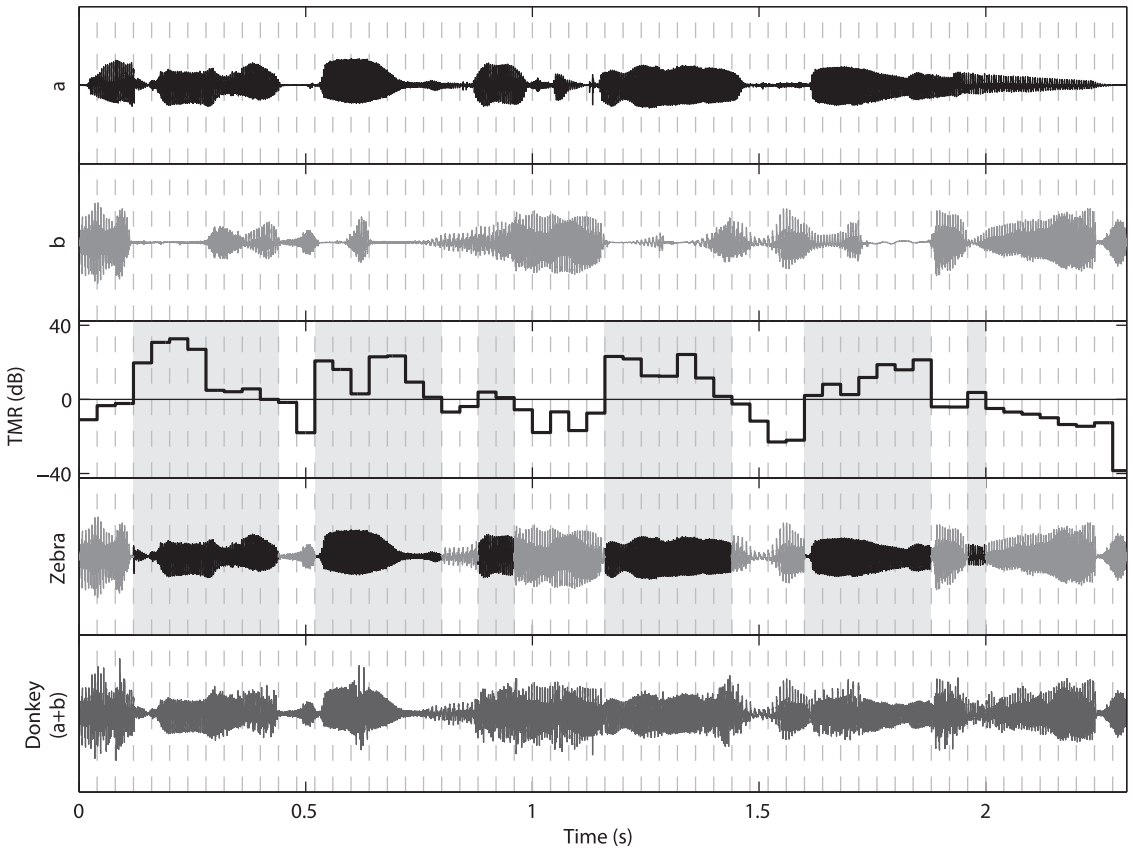


FIG. 2. *First two panels:* Target (a) and masker (b) waveforms with a long-term TMR of 0 dB. The vertical dashed lines show the boundaries of 40 ms chunks. *Third panel:* TMR (20 $\log_{10}$[RMS(a)/RMS(b)]) in each chunk. The darker background identifies chunks where the TMR is positive. *Fourth panel:* Resulting Zebra-speech. *Bottom panel:* Donkey-speech for the same target and masker.

its RMS matched that of the target sentence, i.e., a TMR of 0 dB was used throughout the study.

In these conditions, with chunks of 40 ms, when the male voice was used as a target, it was on average selected 48.9% of the time in the Zebra process (s.d. 5.2%-points, obtained over $N = 1000$ sentences). When the female speaker was used as a target, it was on average selected 55.6% of the time (s.d. 5.4%-points, $N = 1000$). These values are both close to 50%, which would be the expected value if the target and masker voices had identical statistics (such as rise time, number, and duration of pauses, etc.). The average utterance duration for the female talker was 2.30 s (std. 0.10 s) against 2.70 s (std. 0.14 s) for the male talker. Since the two talkers uttered the same sentences, this also means they had different utterance rhythms, which could explain this difference in average selection in the Zebra process. Note that despite these temporal differences the average TMR, i.e., the total energy of the target segments divided by that of the masker segments, was very close to zero both for the female speaker (−0.19 dB) and for the male speaker (+0.04 dB).

#### 2. Noise-band vocoders

Listening through a CI was simulated using a noise band vocoder (Dorman *et al.*, 1997; Shannon *et al.*, 1995). Different vocoders were used in the different experiments presented here but all followed the same general scheme. The input signal was first filtered into different frequency bands. In each band the signal was half-wave rectified and then low-pass filtered (second-order Butterworth filter) to extract the temporal envelope. This envelope was then used to modulate the amplitude of a broadband noise. The resulting signal was then filtered again in the same frequency band, before being added to the output of the other bands. The RMS of the resulting signal was finally adjusted to match that of the input signal.

The frequency bands were extracted using sixth-order Butterworth bandpass filters. Cutoff frequencies for each band varied depending on the number of bands and the overall frequency range covered by the vocoder. Cutoff frequencies for each band of each vocoder used in the experiment are given in Table I. The center frequencies of the bands were chosen so as to be evenly spaced along the cochlea (Greenwood, 1990). In the rest of the article, the vocoders are defined by their overall frequency range between brackets followed by the number of bands, e.g. "[70, 5000] 8-bands."

### B. Procedure

In all experiments the participants were first exposed to some isolated vocoded sentences from the target corpus to familiarize them with the vocoder distortion. In Experiments 1 and 3a, this part consisted of informal listening. The subjects listened to about a dozen vocoded sentences but the

TABLE I. Cutoff frequencies (in Hertz) for the different noise-band vocoders used in the different experiments (bottom row).

| | [70, 5000] 8-bands | | [100, 6000] 8-bands | | [100, 6000] 4-bands | |
|---|---|---|---|---|---|---|
| Band | Lower | Upper | Lower | Upper | Lower | Upper |
| 1 | 70 | 181 | 100 | 228 | 100 | 417 |
| 2 | 181 | 344 | 228 | 417 | 417 | 1114 |
| 3 | 344 | 584 | 417 | 698 | 1114 | 2643 |
| 4 | 584 | 937 | 698 | 1114 | 2643 | 6000 |
| 5 | 937 | 1457 | 1114 | 1730 | | |
| 6 | 1457 | 2221 | 1730 | 2643 | | |
| 7 | 2221 | 3346 | 2643 | 3996 | | |
| 8 | 3346 | 5000 | 3996 | 6000 | | |
| Experiment | 1 | | 3a | | 2a, 2b, 3b | |

exact number varied across participants depending on how easily they felt they could understand vocoded speech. In the other experiments the participants were presented with isolated vocoded sentences from the target corpus while the sentence was written on the screen at the same time. They were instructed to read and listen to the sentences. The exact number of such presentations varied across experiments and is given in the description of each experiment.

After this initial step, in all experiments the participants underwent a training block. In Experiments 1 and 3a, this training block was of 80 trials and contained only Donkey-speech; in the other experiments, the training block had the same number of trials and conditions as the testing block. The participants were instructed to try to extract a color and number from the mixture they heard. They were informed that two competing speakers were mixed together but that only one of them was saying a color and number. A grid containing the eight different numbers in columns repeating in four rows of the four different colors was presented to them on each trial, and they had to click on the cell corresponding to the color and number they heard. Note that although the call sign differed from sentence to sentence, participants did not have to identify it in order to perform the task. In the training trials the response of the listeners was followed by feedback: First the correct response was displayed, blinking, on the screen; then a non-vocoded version of the same mixture of target and masker signals was presented. In all experiments a message was displayed every about 50 trials (unless stated otherwise), prompting the participant to take a short break.

The testing block was similar to the training block except that no feedback was provided. The number of trials in this block is provided in the description of each experiment.

The performance of participants can be scored loosely or strictly. In the loose scoring method participants get a score of 0.5 if they get either the color or the number correct, and 1 if they get both the color and number correct. The strict scoring method consists of giving a score of 1 only when both the color and number are correct, and zero otherwise. The loose scoring method was used throughout the article because it gives better resolution, especially at low scores. The pattern of results was nevertheless generally identical with the strict scoring method. All statistical analyses were performed on scores transformed into rationalized-arcsine units (RAU; Studebaker, 1985). When relevant, repeated-measure analysis of variance (ANOVA) were used. In all ANOVAs the sphericity assumption was tested and always found to be valid, so no correction was applied. Comparisons between individual conditions were tested for significance using *t*-tests. Multiple comparisons were all corrected using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995). Note that only the comparisons reported, rather than all possible comparisons, were considered in the correction, unless stated otherwise. When all conditions were compared to the same control condition, the more appropriate Dunnett test was used instead, using the error term and degrees of freedom of the considered effect in the repeated-measure ANOVA (Dunnett, 1955).

### C. Apparatus

All stimuli in all experiments were presented through an ASUS Xonar Essence STX soundcard, a TDT PA4 attenuator, a TDT HB7 headphone buffer, and Sennheiser HD 650 headphones. The sound level was calibrated to be 75 dB SPL (sound pressure level) in the right ear, for the Donkey condition. This level was measured in the ear canal of a KEMAR Type 45DA head assembly. All the experiments took place in a double-walled sound-treated booth.

## III. EXPERIMENT 1: EFFECT OF CHUNK DURATION

### A. Rationale

Chunk duration is the only parameter in Zebra-speech. The aim of this first experiment was to determine the optimal value needed to maintain intelligibility. Maintaining intelligibility is particularly important because this measure is relevant for ecological situations. It is also important to equate intelligibility between Zebra- and Donkey-speech so that additional manipulations, described in Secs. IV and V, can be compared using a similar baseline and in the absence of floor or ceiling effects.

Given that several consecutive chunks can be sourced from the same speaker, the chunk duration really defines the temporal quantization of the signals. The durations of the segments uttered by a given speaker in the resulting Zebra-speech are integer multiples of the chunk duration. Hence, shorter chunks should provide better precision in the timing of the speaker switch. However, a shorter chunk duration will complicate the extraction of information from single chunks when they occur. Moreover, in order to convey envelope fluctuations corresponding to $F0$, the duration of the chunks needs to be longer than the period imposed by the $F0$ of the speech signal.

In this first experiment, participants had to identify the color and number from vocoded CRM sentences processed using the Donkey or Zebra methods with, in the latter case, different chunk durations. The optimal chunk duration is defined as the longest duration that produced identification scores that were not significantly lower than those obtained with Donkey-speech.

The two speakers, described in Sec. II, were used for the target sentences. Although the two speakers differed in sex and $F0$, the primary goal of using multiple speakers was to generalize the results to different voices rather than to study the specific effect of differences in voice characteristics.

## B. Method

### 1. Stimuli and apparatus

The stimuli were as described in Sec. II A. The Donkey- and Zebra-speech stimuli were processed using the [70, 5000] 8-bands vocoder with an envelope extraction cutoff frequency of 320 Hz. In a control condition, Donkey-speech was also presented unprocessed, i.e., without vocoding, to check that the task was possible. Chunk durations for the Zebra-speech were 10, 20, 40, 80, and 180 ms. Two different speakers, a male and a female, were used for the target speech. The maskers were uttered by another male speaker. All stimuli were presented only in the right ear.

### 2. Participants

Eight native speakers of British English (aged 19 to 30, average 24.5 yrs) were paid to take part. All had NH, i.e., pure tone thresholds of less than 20 dB HL (hearing level) at octave frequencies between 250 and 4000 Hz in the right ear. All listeners gave written informed consent prior to testing and were paid an hourly wage for their participation.

### 3. Procedure

The training block was of 80 trials, all consisting of vocoded mixtures processed using the Donkey method, and including the 32 color-number combinations and the two speakers. The testing block had 40 trials for each speaker in each condition: Non-vocoded Donkey, vocoded Donkey, and vocoded Zebra with 10, 20, 40, 80, and 180 ms chunk duration, yielding a total of 560 trials. A message instructed the participants to have a short break every 80 trials.

## C. Results

Average scores across subjects are shown in Fig. 3. The average score for the non-vocoded (Donkey) condition is greater than 99%, indicating that the participants were able to do the task. Vocoding reduced the average score to 72% and 77% for the male and female speaker, respectively (open symbols). Overall, scores in the Zebra conditions were similar to those in the Donkey condition. A repeated measure ANOVA on the RAU score difference between the Donkey condition and the Zebra conditions using chunk duration and speaker as repeated factors revealed no significant main effect of the speaker [$F(1,7) = 1.19$, $p = 0.31$] but a significant effect of chunk duration [$F(4,28) = 10.06$, $p < 0.001$] and a significant interaction between speaker and chunk duration [$F(4,28) = 4.27$, $p < 0.01$], reflecting the fact that the response pattern for the female speaker had a different shape than the one obtained for the male speaker. Comparisons between scores for each chunk durations and scores obtained in the Donkey conditions are shown for each speaker in

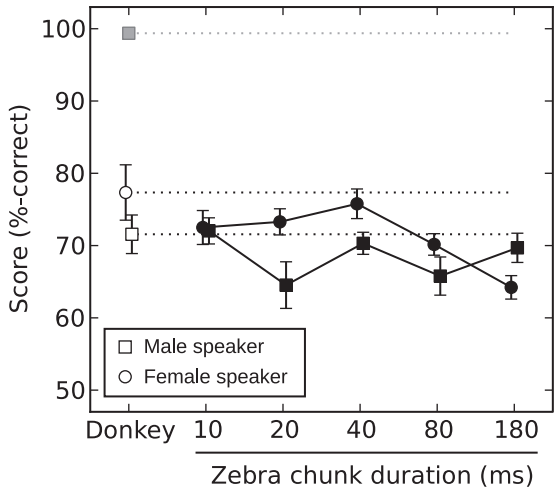


FIG. 3. Experiment 1—Score as a function of condition averaged across listeners. The gray square shows the average score across subjects for unprocessed Donkey-speech. The open symbols show performance for vocoded Donkey-speech. The black symbols show average scores for vocoded Zebra-speech with chunk durations ranging from 10 to 180 ms. Circles represent the female speaker and squares represent the male speaker. Error bars represent the across-subject standard error of the mean.

Table II along with comparisons of the average scores across speakers. The significance of these comparisons was tested using Dunnett's test, and showed that 80 ms was the only chunk duration yielding scores for Zebra-speech that were significantly different from those for Donkey-speech for both speakers as well as for the average.

## D. Discussion

Participants achieved similar scores on average with the "conventional" method (Donkey-speech) as when segments of speech that are likely to be masked were removed and segments that are less masked were made fully available (Zebra-speech). This indicates that the segregation strategy used by the participants in the Donkey condition provides them with an amount of information that is equivalent to the one obtained in the Zebra condition. This conclusion particularly holds for the female speaker for chunk durations up to 40 ms where the differences between Donkey and Zebra

TABLE II. Scores (and RAU scores between brackets) in the Donkey condition subtracted from those in the Zebra condition, Dunnett test $t_d$- and $p$-values for each chunk duration and each speaker. Significant differences are shown in bold type.

| | 10 ms | 20 ms | 40 ms | 80 ms | 180 ms |
|---|---|---|---|---|---|
| Male | −0.5% (−0.4) | 7.0% (6.9) | 1.2% (1.4) | 5.8% (5.7) | 1.9% (2.0) |
| | $t_d = 0.15$ | $t_d = -3.28$ | $t_d = -0.69$ | $t_d = -2.74$ | $t_d = -0.96$ |
| | $p > 0.999$ | $\mathbf{p = 0.010}$ | $p = 0.936$ | $\mathbf{p = 0.039}$ | $p = 0.802$ |
| Female | 4.8% (5.6) | 4.1% (4.9) | 1.6% (2.2) | 7.2% (8.1) | 13.1% (13.9) |
| | $t_d = -2.05$ | $t_d = -1.80$ | $t_d = -0.82$ | $t_d = -2.98$ | $t_d = -5.08$ |
| | $p = 0.172$ | $p = 0.269$ | $p = 0.877$ | $\mathbf{p = 0.022}$ | $\mathbf{p < 0.001}$ |
| Average | 2.2% (2.5) | 5.5% (5.9) | 1.4% (1.7) | 6.5% (6.9) | 7.5% (7.9) |
| | $t_d = -1.64$ | $t_d = -3.92$ | $t_d = -1.15$ | $t_d = -4.54$ | $t_d = -5.24$ |
| | $p = 0.354$ | $\mathbf{p = 0.002}$ | $p = 0.671$ | $\mathbf{p < 0.001}$ | $\mathbf{p < 0.001}$ |

remain below five percentage points. Although the overall difference between Donkey- and Zebra-speech did not significantly depend on the speaker, the effect of chunk duration depended on the speaker. While the pattern of results for the female speaker shows a clear drop in performance after 40 ms, the pattern for the male speaker does not show such a drop. Although the sentences uttered by the male speaker were longer than those uttered by the female speaker, it is unclear how this could yield a non-monotonic result pattern for one speaker and not for the other.

In order to choose the optimal chunk duration, both the average performance and the performance for each speaker were considered. The value of 40 ms is the longest chunk duration yielding performance not significantly different from that for Donkey-speech, for the average across speakers as well as for each of the speakers. This value was therefore used in all subsequent experiments.

In order to compare this 40-ms chunk duration with the literature, it is worth noticing that in the current experiment, the chunk duration is only a quantization measure, i.e., the actual segments consist of a number of consecutive chunks. The average durations of the actual segments were 98, 121, 147, 199, and 310 ms for chunk durations of 10, 20, 40, 80, and 180 ms, respectively. Chunks of 40 ms thus produce segments whose duration (147 ms) roughly corresponds to the syllable rate in British English (e.g., Patel *et al.*, 2006, 5.8 syllables/s, i.e., 172 ms per syllable). For comparison, the keywords of the CRM corpus have an average duration of 330 ms (s.d. 62 ms).

Note that although the principle of Zebra-speech is comparable to that of the interleaved-word procedure introduced by Broadbent (1952) and more recently adapted by Kidd *et al.* (2008), the two methods actually differ by a number of aspects. One key difference is that the interruption pattern in Zebra-speech is based on the acoustical properties of the signals while it is based on the linguistic content in the interleaved-word procedure. As a result, Zebra-speech can operate on much shorter timescales that are more compatible with streaming: In Zebra-speech the average segment duration was 147 ms, while in the interleaved-word procedure the average word duration was 624 ms.

## IV. EXPERIMENT 2: EFFECT OF THE TYPE OF MASKER

When the TMR varies over time, as in Donkey-speech, performance may be limited by several factors: (a) Portions of the target may be absent, or be obliterated by simultaneous energetic masking, (b) portions of the target may be subject to forward masking from earlier portions of the masker, and (c) the listener may not be able to effectively "glimpse" those portions of the mixture where the TMR is highest. This latter factor is likely to be exacerbated both with CI processing and noise-vocode simulations, where differences in spectral and temporal fine structure between the masker and target are substantially degraded (e.g., Qin and Oxenham, 2003; Gnansia *et al.*, 2009; Pierzycki and Seeber, 2010). In Zebra-speech, simultaneous energetic masking is controlled as only one voice is presented at a time but forward masking could still occur as, e.g., reported for interrupted speech by Dirks and Bower (1970). Experiment 2 sought to differentiate between these effects: Experiment 2a focused on the role of target-masker similarity while Experiment 2b specifically addressed the potential role of forward masking in Zebra-speech.

### A. Experiment 2a: Target-masker similarity

#### 1. Methods

Zebra speech was constructed as in the previous experiment but, in different conditions, the chunks where the masker sentence should have been presented were filled with other signals. In the first condition, they were filled with silence (Zebra/silence) and then vocoded. This condition was used to evaluate the effect of removing chunks of the signal in the Zebra process; this also simulates, to some extent, the loss of information produced when portions of the target speech are simultaneously masked by a fluctuating noise masker, although using silence may also introduce erroneous phonological cues. The resulting loss of information can be evaluated by comparing it to the second condition, "Target alone," which consisted simply of a noise-vocoded version of the target in the absence of the masker. The third condition (Zebra/masker) corresponds to the standard Zebra processing as used in Experiment 1. A masker was first created and was then used to compute the Zebra interruption pattern and to fill the gaps created in the target by the Zebra process. The resulting stimulus was then vocoded. In a fourth condition (Zebra/SSN), the gaps were filled with speech-shape noise (SSN), i.e., stationary noise having the same long-term spectrum as the masker sentence. On each of these trials, a speech masker was first created (as in the Zebra/masker condition) and the Zebra interruption pattern was calculated for this masker. The SSN was then derived from this masker by randomizing the phase components of its spectrum. The gaps introduced in the target sentences by the Zebra interruption pattern were then filled with the SSN. The RMS level of the SSN chunks was constant throughout the stimulus and was identical to the overall RMS level of the speech-masker chunks. The resulting signal was then vocoded. The gaps between the target segments were filled with noise, thereby potentially impairing performance both via forward masking and by making the masker difficult to discriminate from the target. To differentiate between these two explanations, a fifth condition (Zebra/SSC) filled the target gaps with a speech-shape complex (SSC) having an $F0$ of 100 Hz. The reasoning was that the SSC would sound qualitatively different from the vocoded speech, and that this difference would reduce the confusion between target and masker, but not affect energetic masking. The components of the SSC were added in cosine phase, and had a spectral envelope matching the long-term spectrum of the masker sentence. The SSC-filled chunks were not passed through the vocoder, and were added to the interrupted and vocoded target signal in order to preserve their harmonic nature. The RMS of the SSC chunks was equal to the RMS of the full masker sentence minus 12 dB. This level adjustment was performed in an attempt to equate both the loudness and the waveform peak amplitude of the SSC and target speech

chunks. In Experiment 2a, the five conditions described above were used along with a Donkey-speech condition.

The [100, 6000] 4-band vocoder was used because if forward masking plays a role, it can be expected to be higher with poorer spectral resolution. Only the female voice ($F3$) from the CRM corpus was used and the masker speech was as described in Sec. II. Four new NH volunteers participated in Experiment 2a.

#### 2. Results and discussion

Results from Experiment 2a are shown in Fig. 4(a). When the full target alone was presented, participants scored more than 88% on average. Removing the chunks with negative TMR (Zebra/silence) reduced the performance to 63%. A similar score (59%) was observed for chunks filled with a SSC. These two conditions (Zebra/silence and Zebra/SSC) were not significantly different from each other [$t(3) = 2.27$, $p = 0.14$] but were both significantly lower than the Target alone condition [$t(3) = 9.19$, $p = 0.007$ and $t(3) = 11.95$, $p = 0.004$, respectively]. The fact that the SSC produced performance that was similar to filling the target gaps with silence suggests that it did not produce substantial forward masking. In contrast, filling the gaps with either the speech masker (Zebra/speech, 43%) or Zebra/SSN (36%) resulted in a significant reduction of performance [compared to the silence-filled gap condition Zebra/silence, $t(3) = 11.29$, $p = 0.005$ and $t(3) = 8.63$, $p = 0.007$, respectively]. Performance in these two conditions was significantly worse than with the SSC [$t(3) = 5.82$, $p = 0.015$ and $t(3) = 5.74$, $p = 0.015$, respectively] but not significantly different from the Donkey condition [$t(3) = -1.78$, $p = 0.19$ and $t(3) = 0.23$, $p = 0.83$, respectively].

These results are consistent with the idea that, in the Zebra speech conditions, the major detriment to performance lies in target-masker discriminability, rather than forward masking. Filling the target gaps with either the vocoded masker or with a steady noise impaired performance but filling it with the SSC did not. However, the SSC used in Experiment 2a may have produced less forward masking than the SSN for two reasons: First, it had a lower RMS level, and second, it had a peaker temporal envelope. Previous studies have shown that temporal properties of a signal influence forward masking; i.e., sounds with identical long-term magnitude spectra can cause different amounts of masking (Carlyon and Datta, 1997; Gockel *et al.*, 2003). In particular, sounds, such as a cosine-phase complex, that produce peaky temporal envelopes at the outputs of peripheral auditory filters can produce less forward masking than sounds with flatter envelopes—for example when excited by a noise, a random phase complex, or a negative Schroeder phase complex.

### B. Experiment 2b: Forward masking

#### 1. Rationale and methods

Experiment 2b controlled for the amount of forward masking by (a) boosting the level of the SSC by 12 dB so that its level was the same as the SSN, and (b) including the SSC-sch$^-$ condition, in which the harmonics were added in "negative Schroeder" phase (Schroeder, 1970), i.e., harmonic $n$ had a phase equal to $-\pi\, n\, (n-1)/N$, where $N$ is the total number of harmonics in the vocoder frequency range. As shown in Fig. 5, this latter condition produces (simulated) auditory filter outputs that are less peaky than for the cosine-phase complex and roughly similar in peakiness to the SSN. The peak factors for the outputs of three auditory filters shown in Fig. 5 are: 8.5, 12.6, and 15.0 dB for SSC-cos; 6.9, 6.8, and 8.2 dB for SSC-sch$^-$; and 7.1, 7.8, and 8.3 dB for SSN. Note that for each filter the peak factors for SSC-sch$^-$ are more similar to those for the SSN than to those for the SSC-cos complex. If forward masking plays a major role, then the results in the SSC-sch$^-$ condition should be more similar to those of the SSN condition than to those of the SSC-cos condition. However, if the role of forward masking is negligible, then performances in the SSC-sch$^-$ condition should resemble those in the SSC-cos condition, and both should be higher than in the SSN condition.

Experiment 2b was therefore identical to Experiment 2a except that there were no "Donkey" or "Target only" conditions, and that different types of SSC were used. Specifically,

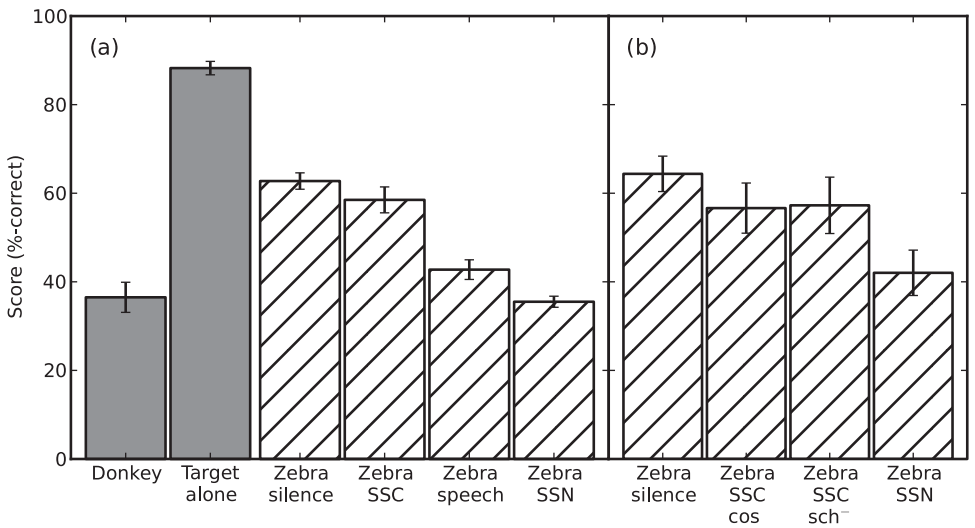


FIG. 4. Experiment 2—Average scores across subject for the Donkey (gray) and Zebra (hatched) conditions. Parts (a) and (b) show data from Experiments 2a and 2b, respectively. Error bars represent the standard error across subjects.

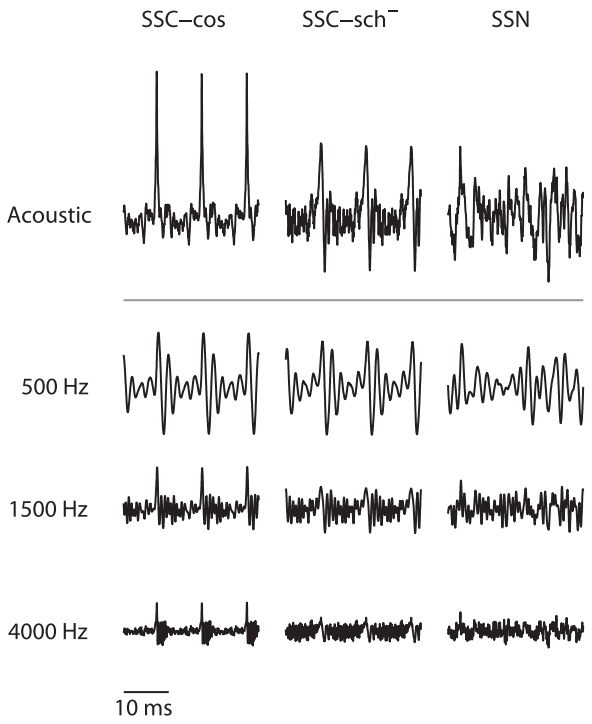


FIG. 5. *Top row*: Acoustic waveforms of the three types of maskers used in Experiment 3. *Bottom rows:* Basilar membrane displacement at three center frequencies; evaluated using a phase-corrected Gamma chirp filterbank (Oxenham and Dau, 2001).

the SSC now had an rms level equal to that of the SSN (i.e., 12 dB higher than in Experiment 2a), and was used in two separate conditions, one with all components added in cosine phase (SSC-cos) and one with components added in negative Schroeder phase (SSC-sch$^-$). One participant of Experiment 2a and seven new listeners took part in Experiment 2b.

#### 2. Results and discussion

As shown in Fig. 4(b), the performance in Experiment 2b was similar for the two SSC complexes, and better than with the SSN. A repeated measure ANOVA on RAU scores with the type of masker as a factor indicates that masker type had a significant effect [$F(3,21) = 13.18$, $p < 0.0001$]. All maskers yielded significantly lower scores than silence [$p = 0.03$ for SSC-cos and $p < 0.01$ for SSN] except SSC-sch$^-$ for which the difference did not reach significance [$p = 0.06$]. The two SSCs were not significantly different [$p = 0.72$] but were both different from the SSN [$p = 0.03$ each]. This suggests that the negative effect of filling the gaps with noise was more likely due to target-masker similarity than to forward masking. A potential *caveat* is that in auditory filters with CFs lower than the lowest CF (500 Hz) shown in Fig. 5, the harmonics of the SSCs would become resolved. The peakiness of the outputs of those filters would not depend on the component phase, and so if intelligibility were dominated by information in these low frequency regions one might not expect to see a phase effect, even if forward masking affected performance. However, it should be noted that a comparison of the results with the cosine-phase SSC complexes in Experiments 2a and 2b—which were identical except for a 12 dB difference in level—revealed performance that was essentially equal in the two conditions. It seems unlikely that this 12 dB difference, present across the whole frequency spectrum, would have a smaller effect on forward masking than would a difference in the peakiness of auditory filter outputs, restricted to low frequencies and with equal RMS.

In conclusion, it seems safe to assume that forward masking plays little to no role in Zebra-speech perception. This means that only across-time linkage of segments, referred to as sequential segregation mechanisms here, are involved in the perception of Zebra-speech, thus validating the method for the study of this particular class of mechanisms.

## V. EXPERIMENT 3: BINAURAL SEGREGATION CUE

### A. Rationale and general method

The main objective of Experiment 3 was to determine whether simulated CI listeners can use binaural configuration for across-time linkage, which we refer to here as sequential segregation. We only considered here the case where the signal in one ear provides information about the signal in the other ear, i.e., no interaural time difference was used and loudness differences were in the form of all or nothing. By comparing how these binaural configurations affect performance for Zebra-speech relative to Donkey-speech, the importance of the sequential mechanism in natural situations can also be estimated.

Using the interleaved-word procedure, Kidd *et al.* (2008) showed that a fixed perceived interaural location contributed to across-time linkage of separated words in NH listeners. However, this finding does not seem to extend to CI users. In a more recent study, Ihlefeld *et al.* (2010) required listeners to identify noise-vocoded speech presented in a 16 Hz square-wave modulated masker. Although performance improved by a contralateral noise that was modulated identically to the masker, this improvement was no greater than that produced by a contralateral steady noise in a condition where the ipsilateral masker was also steady. So in the modulated condition, listeners did not seem to be able to gain a boost in performance from the timing information presented to the contralateral ear, and which might have helped them glimpse unmasked portions of the target. Several explanations may be found for this lack of effect. First, the binaural system is known to be sluggish (see Krumbholz *et al.*, 2009, for a review) and could be unable to provide glimpsing enhancement when the chunks are as short as 32 ms (corresponding to the 16 Hz gating used by Ihlefeld *et al.*, 2010). Indeed, Krumbholz *et al.* (2009) reported signs of binaural sluggishness from rates of 16 Hz and higher. A lower gating frequency, i.e., longer glimpsing opportunities, may help listeners to access binaural cues and thus may improve their glimpsing ability based on such cues. Second, the regularity of the gated noise may have already reduced informational masking or modulation interference. The use of an irregular masker might leave more space for improvement based on binaural cues. Finally, it could simply be that listeners rely more heavily on simultaneous than on sequential grouping cues. By using Zebra speech we were able to

assess the role of binaural cues to sequential grouping under conditions with a longer (147 ms) average segment duration, with an irregularly modulated masker, and where simultaneous grouping cues were not required.

The second aim of the experiment was to study not only sequential segregation on its own but also the relative importance of sequential and simultaneous mechanisms to speech segregation based on a binaural configuration. To do this we presented target-masker mixtures processed using either the Donkey or Zebra method to one ear, and added different types of cue to the other. We have argued above that Zebra-speech requires only sequential segregation, whereas Donkey-speech potentially involves both sequential and simultaneous processes. Our hypothesis is that, if the segregation processes involved in the two types of processing really do differ, then they should respond differentially to the different types of contralateral cue.

The three contralateral cues were "*Full masker*" (a copy of the masker sentence processed in Donkey mode), "*Masker Chunks*" (those portions of the zebra speech mixture corresponding to the masker), and "*Noise Chunks*" (same as Masker Chunks except with a fixed spectrum corresponding the long-term average of the masker). Schematic representations of these conditions are shown in Fig. 6. The stimulus in the left ear, where only the masker was presented, was attenuated by 6 dB to prevent the overall masker from becoming too loud. For the Zebra condition, one would predict that performance would be best when the contralateral cue is presented only during the masker intervals, as this would make the glimpsing strategy easier than when the masker cue is continuous in the contralateral ear. In this condition then, performance should be better with the Masker-Chunks than with the Full-Masker cue. Importantly, the pattern of results predicted for Donkey-speech (which is the mode of processing used in most CI simulations) depends on whether the benefit of a contralateral cue helps primarily sequential or simultaneous segregation. To the extent that it helps sequential segregation then the results should be similar to those with Zebra speech. In contrast, if it mostly helps simultaneous segregation, then performance should be best when the copy of the masker is present at the same time as the target—in other words, it should be better in the Full-Masker than in the Masker-Chunks condition. Finally, in both conditions, performance with the Noise-Chunks contralateral signal should depend on the degree to which the interaural grouping cue used in that condition relies on the spectral similarity between the stimuli in the two ears.

The balance between simultaneous and sequential segregation is also likely to depend on spectral resolution. In the context of simultaneous segregation, a decrease in spectral resolution causes an increase in energetic masking which cannot easily be overcome with a binaural cue. In contrast, although reduced spectral resolution affects the monaural discriminability of successive segments, binaural cues could help restore the ability to identify whether the segments belong to the target or the masker. Therefore, it seems possible that sequential segregation mechanisms become more important as spectral resolution is reduced. This hypothesis was tested by using an 8-band vocoder in Experiment 3a and a 4-band vocoder in Experiment 3b. The general expectation is that the contralateral signal should benefit more the Donkey than the Zebra condition with the higher resolution (8-bands), while the opposite should happen at the lower resolution (4-bands).

In the two experiments, all cue conditions were compared to the reference condition "None," where no sound was presented to the left ear, in order to measure and compare the benefit (or masking release) induced by each cue. The noise carriers in each ear were independent, as in the dichotic condition of Ihlefeld *et al.* (2010). This eliminated any potential benefit due to interaural decorrelation when the masker was presented to one ear. Participants were instructed that the target was always in the right ear alone, and that the masker they had to ignore might be perceived either on the right or in the middle of their head.

## B. Specific methods

### 1. Experiment 3a: 8-bands vocoder

Nine new NH volunteers took part. The [100, 6000] 8-band vocoder was used, and the male (*M*3) and female (*F*3) speakers of the CRM corpus were both used. Donkey- or Zebra-speech was presented to the right ear. For each speaker, each condition was repeated 40 times in random order. Note that, as in Experiment 1, two different speakers were used for sake of generalization over speakers rather than to study the effect of voice differences.

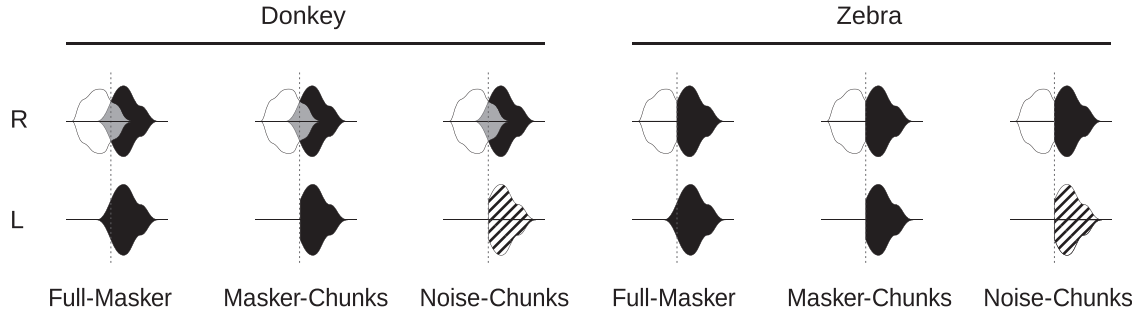


FIG. 6. Schematic representation of the conditions used in Experiments 3a and 3b. The upper row shows stimuli presented to the right ear (R), the bottom row those presented to the left ear (L). The waveform filled in black represents a short segment (a few Zebra-speech chunks) of the masker; the waveform filled in white represents a short segment of the target. Gray areas show the overlap between the two. The signal filled with the hatches represents the SSN filled envelope. The vertical gray dashed line represents the location in time of a switch from target to masker in the Zebra-speech.

### 2. Experiment 3b: 4-bands vocoder

Twelve NH volunteers took part. One had also participated in Experiment 2a, and five had participated in other experiments involving noise-vocoded Zebra-speech. The six remaining subjects were new volunteers. The experimental design was identical to that of Experiment 3a, except that the longer training block was used, only the female talker was used, and that the noise-band vocoder was [100, 6000] 4-bands. Each condition was repeated 50 times in random order.

## C. Results

Average scores for Experiments 3a and 3b are shown in the left-hand panels of Figs. 7(a) and 7(b), respectively. Individual scores were transformed using the rationalized arcsine method (Studebaker, 1985), and then differences between the RAU scores in conditions where a signal was present in the left ear and the reference condition None were computed to obtain a measure of binaural benefit for all the contralateral signals. These benefit scores are shown in the right-hand panels of Fig. 7.

For both parts of the experiment, scores in the None condition were very similar for Donkey and Zebra, and did not differ significantly [$t(8) = -0.02$, $p = 0.99$ with 8-bands; $t(11) = -1.27$, $p = 0.23$ with 4-bands]. Hence, the effect of changing the number of bands (from 8 to 4) was similar for the two types of processing. However, as discussed below, the cues that subjects were using to segregate the target and masker were quite different. This can be most easily illustrated with reference to the benefit scores, shown in the right-hand panels.

### 1. Experiment 3a: 8-bands

For the 8-band vocoder [Fig. 7(a)], all the contralateral signals provided a significant benefit both for Zebra [$t_d > 3.27$, $p < 0.01$ for all contralateral signals, averaged across speakers] and Donkey conditions [$t_d > 3.41$, $p < 0.01$]. The fact that all the binaural cues provided a significant benefit for Zebra-speech indicates that they can be exploited by purely sequential mechanisms.

The amount of benefit depended both on the nature of the contralateral signal [$F(2,16) = 11.77$, $p < 0.01$] and on whether Donkey or Zebra was presented in the attended ear [$F(1,8) = 9.68$, $p = 0.014$]. Importantly, there was a highly significant interaction between the type of processing and the nature of the contralateral cue [$F(2,16) = 89.79$, $p < 0.0001$] indicating that the different cues did not affect Zebra- and Donkey-speech perception in the same way. Benefits in all conditions were then compared to each other using FDR corrected paired $t$-tests. The results of the Zebra-speech conditions are first examined in order to determine the factors affecting sequential segregation. These results are then compared to those obtained for Donkey-speech to evaluate the relative role of sequential and simultaneous segregation.

*a. Zebra-speech:* Sequential segregation. As predicted, for Zebra-speech performance was better with the Masker-Chunks than with the Full-Masker cue [$p < 0.001$]. In the Zebra/Full-Masker, the temporal envelope of the masker signal in the left ear was continuous whereas in the attended

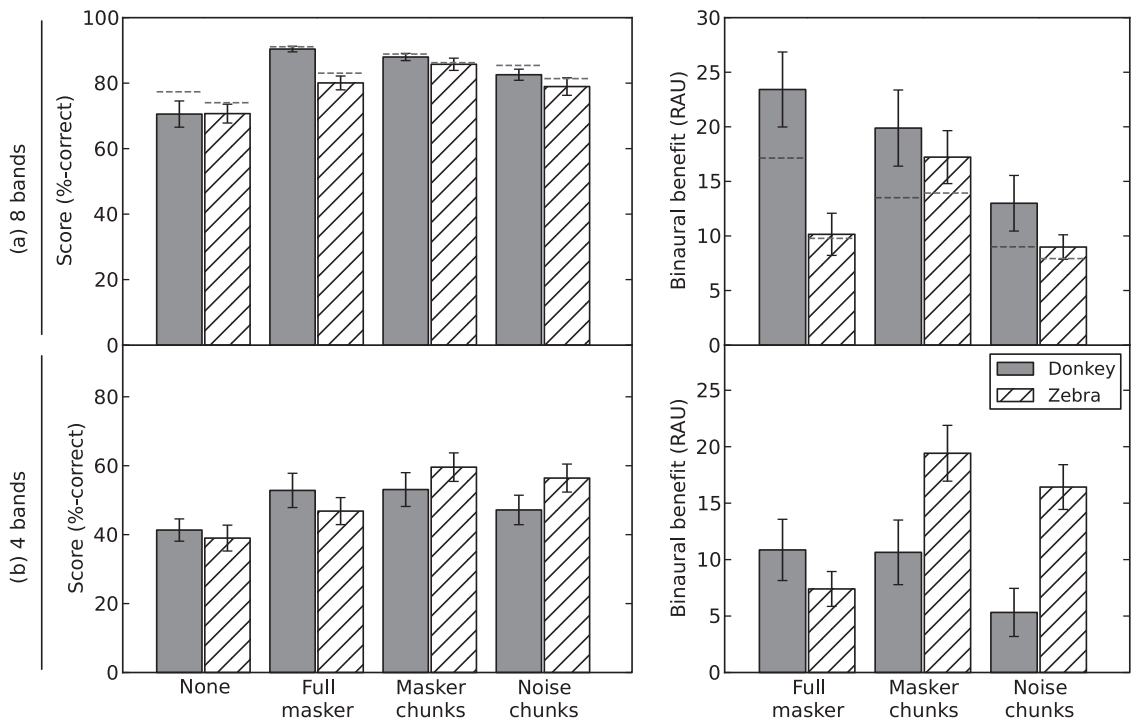


FIG. 7. Results for Experiments 3a (top, 8-bands) and 3b (bottom, 4-bands). *Left panels:* Average scores across subjects and speakers for Donkey- (gray) and Zebra-speech (hatched) for all the binaural conditions. *Right panels:* Average binaural benefit in RAU scores for Donkey- and Zebra-speech for the three conditions presenting a signal in the left ear, relative to the None condition. The dashed dark gray lines in the top row represent the scores and benefits for the female speaker only.

ear, it was interrupted by the target signal. To perceive a binaural cue, i.e., to perceive the masker in a different location, the signals in the two ears must be perceptually bound together. If the binding fails, then three unrelated sources may be heard: The target and masker in the right ear, and an additional unrelated signal in the left ear. Because this signal is perceived as unrelated, it does not inform the listener about the masker in the right ear and does not produce any masking release. The temporal mismatch between the two ears could have hindered this binaural binding of the two signals representing the masker, thus making the identification of target segments more difficult in the Zebra/Full-Masker condition than in the Zebra/Masker-Chunks condition. This would be consistent with the fact that Noise-Chunks provided a significant benefit for Zebra-speech although the noise chunks only provide a purely temporal cue indicating when not to listen. The fact that the benefit provided by this contralateral signal was smaller than the one provided by Masker-Chunks [$p < 0.01$] suggests that spectral matching across ears is also important for binaural grouping.

*b. Donkey-speech:* Simultaneous and sequential segregation. For Donkey-speech, the Full-Masker cue provided a larger benefit than the Masker-Chunks [$p = 0.037$], i.e., the opposite result pattern to that for Zebra-speech [$t(8) = 11.82$, $p < 0.001$ for the interaction]. This is consistent with the binaural-binding hypothesis since the interaural temporal match was strong in the Donkey/Full-Masker condition whereas it was weaker in the Donkey/Masker-Chunks conditions (see Fig. 5). However, although only sequential segregation was involved in the Zebra conditions, both sequential and simultaneous segregation were involved in the Donkey conditions and binaural binding can strengthen both of these mechanisms. In comparing the two conditions where the binaural binding was strong, it appears that the benefit in the Donkey/Full-Masker condition was larger than the one observed in the Zebra/Masker-Chunks condition [$p = 0.037$]. This indicates that in the Donkey/Full-Masker condition, listeners were able to retrieve some information from the segments where the TMR was negative by using simultaneous segregation.

To assess whether sequential segregation was improved at all in the Donkey conditions, one can note that a small but significant benefit was observed for Donkey-speech even in the Noise-Chunks condition. This condition provided very limited information on the instantaneous masker spectrum, and so arguably provided essentially a temporal glimpsing cue indicating only when to listen. Hence, it appears that the Full-Masker likely improved the perception of Donkey-speech by enhancing both the sequential and the simultaneous segregation, while the Masker-Chunks improved the perception of Zebra-speech by enhancing only the sequential segregation.

Finally, the speaker also had a general effect on performance [the female speaker yielded 83%-correct against 78% for the male speaker, $F(1,8) = 11.54$, $p < 0.01$] but did not interact with any of the other factors. Because no interaction was observed, it was assumed that the only effect the speaker had was to vary the overall intelligibility. For this reason, only the female speaker, who was the most intelligible, was used in Experiment 3b.

### 2. Experiment 3b: 4-bands

As with 8-bands, all contralateral signals provided a significant benefit [$t_d > 4.00$, $p < 0.001$ for all conditions except Donkey/Noise-Chunks: $t_d = 2.23$, $p = 0.08$]. A repeated measures ANOVA on the binaural benefit showed that although on average Zebra-speech was not significantly different from Donkey-speech [$F(1,11) = 3.66$, $p = 0.08$], the effects of the contralateral signals were different from each other [$F(2,22) = 10.01$, $p < 0.01$]. Again, there was a significant interaction between the effects of the two speech processing methods and the contralateral signals [$F(2,22) = 20.48$, $p < 0.001$]. For the Zebra speech, as predicted, performance was once more superior with the masker chunks than with the full masker [$p < 0.01$]. This difference was not observed for Donkey–speech [$p = 0.93$], although, unlike with 8-bands, the opposite effect (better performance with Full-Masker) was not observed either. A combined analysis was conducted to focus on the differences between the two experiments, i.e., how spectral resolution affected sequential segregation and the balance between the two segregation mechanisms.

### 3. Effect of spectral resolution

Binaural benefits (in RAU) from the two experiments were analyzed in a mixed repeated-measure ANOVA with the speech-processing method and contralateral signal as within-subject repeated factors, and the number of bands as the between-subject factor. Only the scores for the female speaker were considered for Experiment 3a to have the same speaker in the two experiments. The scores for the female speaker are shown by dashed lines in Fig. 7(a).

The combined analysis, not surprisingly, confirmed the main trends present in the data from the individual experiments. The nature of the contralateral signal had a significant main effect [$F(2,38) = 11.95$, $p < 0.001$] and interacted with the speech processing method [$F(2,38) = 26.64$, $p < 0.001$].

The nature of the contralateral signal also interacted with the number of bands [$F(2,38) = 6.84$, $p < 0.01$]. Finally, the three-way interaction between processing method, cue, and the number of bands was also significant [$F(2,38) = 3.28$, $p = 0.048$]. No overall effect was found for the speech processing method [$F(1,19) = 0.46$, $p = 0.49$] or for the number of bands [$F(1,19) = 0.004$, $p = 0.95$]; the interaction between the two just missed significance [$F(1,19) = 4.09$, $p = 0.057$].

The interaction between the number of bands and the contralateral cue reflects a finding apparent in Fig. 7 and confirmed by FDR-corrected contrasts. In the Zebra condition, the Noise-Chunks produced better performance than the Full-Masker with 4-bands, but not with 8-bands [$t(18.85) = -5.03$, $p < 0.001$]. The relative performance in these two conditions probably reflects a trade-off between the better temporal cueing in the Noise-Chunks condition and the better spectral match between the two ears in the Full-Masker condition. This would be consistent with the idea that the interaural spectral mismatch in the Noise-Chunks condition was smaller with fewer bands, leading to better performance with the Noise-Chunks than for the Full-Masker cue.

The three-way interaction between the number of bands, contralateral cue, and signal processing method also reflects

a finding apparent in Fig. 7 and confirmed by FDR-corrected contrasts. With 8-bands, performance is better with Donkey/Full-Masker than with Zebra/Masker-Chunks while the opposite pattern is observed with 4-bands [$t(18.50) = 2.32$, $p = 0.032$]. The Zebra/Masker-Chunks condition is the one in which the most effective sequential grouping cue—and only that cue—is expected to influence performance. In contrast, the Donkey/Full-Masker condition is the one in which simultaneous grouping cues should be strongest, even though some glimpsing information arises from the fact that the level of the Full-Masker in the left ear correlates negatively with the instantaneous TMR of the Donkey-speech presented in the right ear. The fact that the relative performance in these two conditions depends on the number of bands suggests that the balance between sequential and simultaneous grouping cues depends on the amount of spectral detail in the stimulus, with simultaneous cues depending more strongly than sequential cues on spectral detail. This too is intuitively plausible, since simultaneous separation must surely depend on the masker having the same spectrum in the two ears, whereas a glimpsing strategy could, at least in principle, be conveyed by any stimulus whose envelope is correlated with that of the masker.

### D. Discussion

All contralateral signals provided a significant benefit for Zebra-speech perception, indicating that sequential segregation, or across-time linkage, can be driven by binaural cues. Both temporal and spectral matching across the two ears were found to improve the binaural binding, which in turn strengthened the segregation. The same factors were found to affect the perception of Donkey-speech, which involves both sequential and simultaneous segregation. The relative contribution of simultaneous and sequential segregation seemed to depend on frequency resolution, with sequential segregation becoming more predominant at poorer frequency resolutions.

Our results show a benefit from temporal matching of the envelopes, rather than the fine structure, of the stimuli presented to the two ears. This contrasts with the results of Ihlefeld *et al.* (2010), who only observed binaural benefits when the noise was correlated across the two ears, and who concluded that the binaural advantage could not be based on the temporal envelope matching. So while these authors interpreted their binaural benefit as due to binaural decorrelation processing (e.g., Edmonds and Culling, 2005) based on the fine structure of the signal, our binaural benefit was likely due to interaural level differences evaluated on the temporal envelope and spectral content. The use of those differences depended on the spectral match between the two ears but could still be obtained when this match was degraded, as in the noise chunks condition. The discrepancy between our results and those of Ihlefeld *et al.* could be due to the fact that our stimuli had more complex temporal envelopes and more spectral detail than the SSN used in their experiment. Spectral binding might be easier when distinctive features can be identified in the spectrum. The temporal envelope they used for their masker was also very simple as it consisted of a regular square wave modulation. In this situation it is possible that the entire benefit was already achieved monaurally by exploiting the regularity of the masker. Adding a contralateral copy of the masker would not yield an additional binaural benefit in that case. When the masker modulation is irregular, as was the case in our experiment, there is a better opportunity for improving performance when a cue describing the masker alone is added. The finding that one potential benefit of contralateral cues is to help the listener glimpse a target in the presence of an irregularly modulated masker has a potentially important practical implication. Tests that measure the benefit of providing unilaterally implanted CI users with a hearing aid or CI in the other ear may underestimate this benefit if they measure speech perception in the presence of unmodulated or regularly modulated maskers.

## VI. GENERAL DISCUSSION

### A. Zebra-speech as a tool to study sequential segregation in concurrent speech perception

While many studies contrast energetic and informational masking in their rationale, some suggest that when energetic masking is accounted for, performance is limited by sequential mechanisms, such as "auditory fusion across temporal gaps" (Nelson and Jin, 2004), or "across-time linkage of target segments (streaming)" (Ihlefeld and Shinn-Cunningham, 2008). In the current experiment, we used noise-vocoded Zebra-speech to directly study the factors affecting the sequential aspect of concurrent speaker segregation in CI simulations.

In Experiment 2, chunks with a positive TMR were replaced by the target alone (mimicking perfect simultaneous segregation in these segments), while chunks with negative TMR were replaced by silence, by a SSN, by a SSC, or by the masker alone. Performance was similar when the gaps were filled with silence or the SSC, and substantially better than when filled by either the SSN or the masker sentence. Performance did not depend substantially either on the level of the SSC or on the phase relationship between its harmonics. Taken together, these findings show that the amount of "sequential interference"—or more exactly the failure to stream segregate—is greatest when there is no qualitative difference between the target and interfering sounds, and, at least for the stimuli used here, is not strongly influenced by forward masking.

The above conclusion is consistent with previous studies on sequential segregation using non-speech sounds that were summarized by Moore and Gockel (2002) by the sentence: "*Any* sufficiently salient perceptual difference may lead to stream segregation…." The present results thus extend previous findings on sequential segregation to natural speech situations. It is worth noticing that unlike the artificial steady-state sounds used in the studies reviewed by Moore and Gockel (2002), the perceptual difference in speech sounds concerns features that can be considered relatively constant or slowly changing while the actual signal and its phonological content is rapidly changing. Furthermore, although their review concerns primarily primitive segregation cues, we have generally interpreted our results in terms

of glimpsing, which is a more general concept. Note also that our results are consistent with previous assertions (e.g., Lorenzi *et al.*, 2006) that spectral or temporal fine structure is important for sound segregation but show that fine structure helps even when it does not convey any information about the target. We should stress though that the issue of whether it is the temporal or spectral fine structure that is important remains unanswered.

Finally, in situations where simultaneous and sequential segregation are non-trivial, the target signal extracted by the listener might be incomplete or contain misleading information. To infer the meaning of the target signal from this degraded input, a third class of mechanisms is required, often referred to as *phonemic restoration* (Powers and Wilcox, 1977; e.g., Warren, 1970) although it actually also encompasses inferences made at the lexical or semantic levels. Recent studies have suggested that poor spectral resolution in CI listeners could also hinder this mechanism (Başkent, 2012; Chatterjee *et al.*, 2010). There is a possibility that this could, in turn, make simultaneous and sequential segregation more difficult, thus creating a vicious circle which would further contribute to degrade intelligibility.

### B. Potential practical applications of Zebra-speech

Because sequential segregation is not limited by energetic masking, it is more versatile than simultaneous segregation in the type of cue that can be used, and has therefore a greater potential for improvement than the latter. This can be illustrated by some of the results of Experiment 2a where the Zebra/SSC condition produced an average score 22 percentage-points (20 RAU) above that of the Donkey condition [$t(3) = 9.99$, $p < 0.01$]. In the Zebra/SSC condition, segments where the TMR is positive are replaced with the target voice only (the TMR becomes $+\infty$) while the segments with a negative TMR are replaced with a very distinctive complex sound, thus reducing confusion between the target and masker to its minimum.

In practice, if perceived differences between the attended voices could be enhanced, thus restoring the saliency of the perceptual difference between the voices, sequential segregation could improve and intelligibility would increase. This is true both for Donkey- and Zebra-speech but it might only be practically feasible for Zebra-speech. In our experiments the Zebra-speech was not constructed from the Donkey-speech signals. Instead they were both created from two separate signals, one considered the target, the other being the masker. Identification of *both* sources to enhance their separability before mixing back presents a non-trivial problem to realistic signal-processing algorithms. However, in the Zebra approach, for any one "chunk," only the more intense source—i.e., the easiest to extract from the mixture—is presented. In each chunk, the dominant signal could then be extracted and modified in order to enhance a specific property ($F0$, vocal-tract length,…), and then directly presented to the listener.

This approach also potentially overcomes a general problem with systems aiming to improve speech perception in adverse environments, which is that they try to suppress the masking sound. For instance, noise-reduction algorithms remove spectro-temporal cells in which noise is detected (Wang and Brown, 2006). Directional microphones also remove a part of the signal that is supposed to be irrelevant for the listener. In both methods an assumption is made about which signal is relevant. However, in the situation where two speakers are competing, no prior assumption can be made about which of the signals is the target and which is the masker and should be suppressed. The target and masker might actually swap roles as the listener shifts their attention from one speaker to the other. Zebra-speech processing does not make any assumptions about which stimulus is the target signal, and presents information from all sources in an interleaved manner.

There is nevertheless a cost, which is that portions of the target speech that are at a negative TMR are removed completely. A benefit from Zebra-speech is therefore likely to be achieved only in listeners who are particularly poor at simultaneous segregation at negative TMRs. The effect of the number of bands on the results of Experiment 3 suggests that patients with poor spectral resolution could directly benefit from Zebra-speech. This is supported by the fact that, in this experiment, the highest score with the 4-band noise-vocoder was obtained with Zebra-speech (absolute score in the Zebra/Masker-Chunks condition greater than every other conditions $p < 0.03$, except Zebra/Noise-Chunks, $p = 0.32$; using Dunnet's test). Note, however, that it has been argued that CI users can show longer recovery times for forward masking (Nelson and Donaldson, 2001). Although Experiment 2b showed that forward masking is not involved in Zebra-speech perception in CI simulations, longer recovery times in real CI users could limit the potential benefit of the method.

## VII. CONCLUSIONS

(1) A new processing method called Zebra-speech has been introduced to investigate sequential aspects—across-time linkage—of concurrent speech segregation.
(2) Sequential mechanisms are largely involved in concurrent speaker segregation, and are driven by the salience of the segregation cue. Forward masking has, at most, a minor role in Zebra-speech perception.
(3) Even when Zebra speech produces the same overall level of performance as with conventional vocoding methods, the segregation cues involved are quite different, as evidenced by differential sensitivity to different types of contralateral cue.
(4) Binaural cues can provide a segregation benefit in CI simulation, even when the fine structure is uncorrelated between the two ears, and sequential segregation plays a significant role in this benefit.
(5) Both sequential and simultaneous segregation are involved in the separation of two concurrent talkers. Our data are consistent with the idea that simultaneous segregation is more strongly affected by reduced spectral resolution than sequential segregation.

## ACKNOWLEDGMENT

This work was performed as part of UK Medical Research Council programme MC_A060-5PQ70.

Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.
Başkent, D. (**2012**). "Effect of speech degradation on top-down repair: Phonemic restoration with simulations of cochlear implants and combined electric-acoustic stimulation," J. Assoc. Res. Otolaryngol. **13**, 683–692.
Benjamini, Y., and Hochberg, Y. (**1995**). "Controlling the false discovery rate: A practical and powerful approach to multiple testing," J. R. Stat. Soc. Ser. B (Stat. Methodol.) **57**, 289–300.
Bernstein, J. G. W., and Brungart, D. S. (**2011**). "Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio," J. Acoust. Soc. Am. **130**, 473–488.
Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.
Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA), 773 p.
Broadbent, D. E. (**1952**). "Failures of attention in selective listening," J. Exp. Psychol. **44**, 428–433.
Buus, S. (**1985**). "Release from masking caused by envelope fluctuations," J. Acoust. Soc. Am. **78**, 1958–1965.
Carlyon, R. P., and Datta, A. J. (**1997**). "Excitation produced by Schroeder-phase complexes: Evidence for fast-acting compression in the auditory system," J. Acoust. Soc. Am. **101**, 3636–3647.
Carlyon, R. P., Long, C. J., Deeks, J. M., and McKay, C. M. (**2007**). "Concurrent sound segregation in electric and acoustic hearing," J. Assoc. Res. Otolaryngol. **8**, 119–133.
Chatterjee, M., Peredo, F., Nelson, D., and Başkent, D. (**2010**). "Recognition of interrupted sentences under conditions of spectral degradation," J. Acoust. Soc. Am. **127**, EL37–EL41.
Chatterjee, M., Sarampalis, A., and Oba, S. I. (**2006**). "Auditory stream segregation with cochlear implants: A preliminary report," Hear. Res. **222**, 100–107.
Cooke, M. (**2003**). "Glimpsing speech," J. Phonetics **31**, 579–584.
Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.
Cooper, H. R., and Roberts, B. (**2009**). "Auditory stream segregation in cochlear implant listeners: Measures based on temporal discrimination and interleaved melody recognition," J. Acoust. Soc. Am. **126**, 1975–1987.
Culling, J. F., and Darwin, C. J. (**1993**). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," J. Acoust. Soc. Am. **93**, 3454–3467.
de Cheveigné, A. (**1999**). "Waveform interactions and the segregation of concurrent vowels," J. Acoust. Soc. Am. **106**, 2959–2972.
Devergie, A., Grimault, N., Tillmann, B., and Berthommier, F. (**2010**). "Effect of rhythmic attention on the segregation of interleaved melodies," J. Acoust. Soc. Am. **128**, EL1–EL7.
Dirks, D. D., and Bower, D. (**1970**). "Effect of forward and backward masking on speech intelligibility," J. Acoust. Soc. Am. **47**, 1003–1008.
Dorman, M. F., Cutting, J. E., and Raphael, L. J. (**1975**). "Perception of temporal order in vowel sequences with and without formant transitions," J. Exp. Psychol. Hum. Percept. Perform. **104**, 147–153.
Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.
Dunnett, C. (**1955**). "A multiple comparison procedure for comparing several treatments with a control," J. Am. Stat. Assoc. **50**, 1096–1121.
Edmonds, B. A., and Culling, J. F. (**2005**). "The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay," J. Acoust. Soc. Am. **117**, 3069–3078.
Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2007**). "Effect of spectral smearing on the perceptual segregation of vowel sequences," Hear. Res. **231**, 32–41.
Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2008**). "Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation," J. Acoust. Soc. Am. **124**, 3076–3087.
Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2012**). "The relationship between concurrent speech segregation, pitch-based streaming of vowel sequences, and frequency selectivity," Acta. Acust. Acust. **98**, 317–327.
Glasberg, B. R., and Moore, B. C. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Gnansia, D., Péan, V., Meyer, B., and Lorenzi, C. (**2009**). "Effects of spectral smearing and temporal fine structure degradation on speech masking release," J. Acoust. Soc. Am. **125**, 4023–4033.
Gnansia, D., Pressnitzer, D., Péan, V., Meyer, B., and Lorenzi, C. (**2010**). "Intelligibility of interrupted and interleaved speech for normal-hearing listeners and cochlear implantees," Hear. Res. **265**, 46–53.
Gockel, H., Moore, B. C. J., Patterson, R. D., and Meddis, R. (**2003**). "Louder sounds can produce less forward masking: Effects of component phase in complex tones," J. Acoust. Soc. Am. **114**, 978–990.
Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Hong, R. S., and Turner, C. W. (**2006**). "Pure-tone auditory stream segregation and speech perception in noise in cochlear implant recipients," J. Acoust. Soc. Am. **120**, 360–374.
Hong, R. S., and Turner, C. W. (**2009**). "Sequential stream segregation using temporal periodicity cues in cochlear implant recipients," J. Acoust. Soc. Am. **126**, 291–299.
Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**, 2915–2922.
Ihlefeld, A., Deeks, J. M., Axon, P. R., and Carlyon, R. P. (**2010**). "Simulations of cochlear-implant speech perception in modulated and unmodulated noise," J. Acoust. Soc. Am. **128**, 870–880.
Ihlefeld, A., and Shinn-Cunningham, B. (**2008**). "Spatial release from energetic and informational masking in a selective speech identification task," J. Acoust. Soc. Am. **123**, 4369–4379.
Kidd, G., Jr., Best, V., and Mason, C. R. (**2008**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**, 3793–3802.
Kitterick, P. T., Bailey, P. J., and Summerfield, A. Q. (**2010**). "Benefits of knowing who, where, and when in multi-talker listening," J. Acoust. Soc. Am. **127**, 2498–2508.
Krumbholz, K., Magezi, D. A., Moore, R. C., and Patterson, R. D. (**2009**). "Binaural sluggishness precludes temporal pitch processing based on envelope cues in conditions of binaural unmasking," J. Acoust. Soc. Am. **125**, 1067–1074.
Kwon, B. J., Perry, T. T., Wilhelm, C. L., and Healy, E. W. (**2012**). "Sentence recognition in noise promoting or suppressing masking release by normal-hearing and cochlear-implant listeners," J. Acoust. Soc. Am. **131**, 3111–3119.
Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.
Luo, X., and Fu, Q.-J. (**2009**). "Concurrent-vowel and tone recognitions in acoustic and simulated electric hearing," J. Acoust. Soc. Am. **125**, 3223–3233.
Luo, X., Fu, Q.-J., Wu, H.-P., and Hsu, C.-J. (**2009**). "Concurrent-vowel and tone recognition by Mandarin-speaking cochlear implant users," Hear. Res. **256**, 75–84.
Mackersie, C. L., Prida, T. L., and Stiles, D. (**2001**). "The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss," J. Speech Lang. Hear. Res. **44**, 19–28.
MacLeod, A., and Summerfield, Q. (**1987**). "Quantifying the contribution of vision to speech perception in noise," Br. J. Audiol. **21**, 131–141.
Mesgarani, N., and Chang, E. F. (**2012**). "Selective cortical representation of attended speaker in multi-talker speech perception," Nature **485**, 233–236.
Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.
Moore, B. C. J., and Gockel, H. (**2002**). "Factors influencing sequential stream segregation," Acta. Acust. Acust. **88**, 320–333.
Nelson, D. A., and Donaldson, G. S. (**2001**). "Psychophysical recovery from single-pulse forward masking in electric hearing," J. Acoust. Soc. Am. **109**, 2921–2933.
Nelson, P. B., and 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., and Nelson, D. A. (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **113**, 961–968.
Nooteboom, S. G., Brokx, J. P. L., and de Rooij, J. J. (**1978**). "Contributions of prosody to speech perception," in *Studies in the Perception of Language*, edited by W. J. M. Levelt and G. B. F. d'Arcais (Wiley and Sons, New York), pp. 75–107.

Oxenham, A. J., and Dau, T. (**2001**). "Reconciling frequency selectivity and phase effects in masking," J. Acoust. Soc. Am. **110**, 1525–1538.
Patel, A. D., Iversen, J. R., and Rosenberg, J. C. (**2006**). "Comparing the rhythm and melody of speech and music: The case of British English and French," J. Acoust. Soc. Am. **119**, 3034–3047.
Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli," J. Acoust. Soc. Am. **59**, 640–654.
Pierzycki, R. H., and Seeber, B. U. (**2010**). "Indications for temporal fine structure contribution to co-modulation masking release," J. Acoust. Soc. Am. **128**, 3614–3624.
Powers, G. L., and Wilcox, J. C. (**1977**). "Intelligibility of temporally interrupted speech with and without intervening noise," J. Acoust. Soc. Am. **61**, 195–199.
Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.
Qin, M. K., and Oxenham, A. J. (**2005**). "Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification," Ear Hear. **26**, 451–460.
Scheffers, M. T. M. (**1983**). "Sifting vowels. Auditory pitch analysis and sound segregation," Ph.D. thesis, University of Groningen, The Netherlands.
Schroeder, M. (**1970**). "Synthesis of low-peak-factor signals and binary sequences with low autocorrelation," IEEE Trans. Inf. Theory **16**, 85–89.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Summers, V., and Leek, M. R. (**1998**). "FO processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss," J. Speech Lang. Hear. Res. **41**, 1294–1306.
van Noorden, L. P. A. S. (**1975**). "Temporal coherence in the perception of tones sequences," Ph.D. thesis, Eindhoven University of Technology, The Netherlands.
Vestergaard, M. D., Fyson, N. R. C., and Patterson, R. D. (**2009**). "The interaction of vocal characteristics and audibility in the recognition of concurrent syllables," J. Acoust. Soc. Am. **125**, 1114–1124.
Vestergaard, M. D., and Patterson, R. D. (**2009**). "Effects of voicing in the recognition of concurrent syllables," J. Acoust. Soc. Am. **126**, 2860–2863.
Wang, D., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (Wiley Interscience, Hoboken, NJ), 432 p.
Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.