Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation

Etienne Gaudrain, Nicolas Grimault, Eric W. Healy, Jean-Christophe Bera

Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation[a)]

Etienne Gaudrain[b)] and Nicolas Grimault[c)]
Neurosciences Sensorielles, Comportement, Cognition, CNRS UMR 5020, Université Lyon 1, 50 Avenue Tony Garnier, 69366 Lyon Cedex 07, France

Eric W. Healy[d)]
Speech Psychoacoustics Laboratory, Department of Communication Sciences and Disorders, University of South Carolina, Columbia, South Carolina 29208

Jean-Christophe Béra
Inserm U556, 151, cours Albert Thomas, 69424 Lyon Cedex 03, France

(Received 17 December 2007; revised 21 August 2008; accepted 22 August 2008)

Cochlear-implant (CI) users often have difficulties perceiving speech in noisy environments. Although this problem likely involves auditory scene analysis, few studies have examined sequential segregation in CI listening situations. The present study aims to assess the possible role of fundamental frequency (F_0) cues for the segregation of vowel sequences, using a noise-excited envelope vocoder that simulates certain aspects of CI stimulation. Obligatory streaming was evaluated using an order-naming task in two experiments involving normal-hearing subjects. In the first experiment, it was found that streaming did not occur based on F_0 cues when natural-duration vowels were processed to reduce spectral cues using the vocoder. In the second experiment, shorter duration vowels were used to enhance streaming. Under these conditions, F_0-related streaming appeared even when vowels were processed to reduce spectral cues. However, the observed segregation could not be convincingly attributed to temporal periodicity cues. A subsequent analysis of the stimuli revealed that an F_0-related spectral cue could have elicited the observed segregation. Thus, streaming under conditions of severely reduced spectral cues, such as those associated with CIs, may potentially occur as a result of this particular cue. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2988289]

PACS number(s): 43.66.Mk, 43.66.Sr, 43.71.Es, 43.71.Ky [JCM] Pages: 3076–3087

I. INTRODUCTION

The mechanisms involved in auditory stream segregation have been thoroughly investigated in normal-hearing (NH) listeners (e.g., Bregman and Campbell, 1971; van Noorden, 1975; Bregman, 1990). These studies led to the *peripheral channeling theory* (Hartmann and Johnson, 1991), which states that two stimuli need to excite different peripheral neural populations to produce auditory streaming. This theory and its implementations (Beauvois and Meddis, 1996; McCabe and Denham, 1997) assume that the main cues for streaming are spectral, suggesting that frequency selectivity is critical. Moore and Gockel (2002), in a review of studies involving sequential stream segregation, further concluded that *any* sufficiently salient perceptual difference may lead to stream segregation, regardless of whether or not it involves peripheral channeling (see also Elhilali and Shamma, 2007). Frequency selectivity can also affect the perceptual salience of cues, and difference limen (DL) measurements can be used to evaluate the salience of stimuli along a given perceptual dimension. Rose and Moore (2005) tested this hypothesis and found that the fission boundary (cf. van Noorden, 1975) was indeed proportional to the frequency DL for pure tones between 250 and 2000 Hz. However, it can be difficult to clearly define the salience of complex sounds composed of many interacting features. Moreover, this difficulty can be compounded when the signal is degraded by the hearing system, such as in hearing-impaired (HI) listeners or in cochlear-implant (CI) users. The current study aims to clarify the role of fundamental frequency in the perceptual segregation of vowel sequences having spectral cues reduced through the use of an acoustic vocoder model of a CI (cf. Dudley, 1939; Shannon *et al.*, 1995).

Experiments involving NH listeners have shed light on the mechanisms underlying pitch-based streaming and on the influence of reduced frequency resolution. Streaming based on fundamental frequency (F_0) is reduced when the resolvability of harmonic components of complex tones is reduced, but it is still possible to some extent even when harmonics are totally unresolved (Vliegen and Oxenham, 1999; Vliegen *et al.*, 1999; Grimault *et al.*, 2000). Gaudrain *et al.* (2007)

[a)]Portions of this work were presented in "Segregation of vowel sequences having spectral cues reduced using a noise-band vocoder." Poster presented at the 151st ASA meeting in Providence, RI, June 2006.
[b)]Present address: Centre for the Neural Basis of Hearing, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, United Kingdom.
[c)]Electronic mail: ngrimault@olfac.univ-lyon1.fr
[d)]Present address: Department of Speech and Hearing Science, The Ohio State University, Columbus, 43210.

found that F_0-based streaming of vowel sequences was reduced when frequency resolution was reduced by simulated broad auditory filters (Baer and Moore, 1993). Roberts *et al.* (2002) showed that differences solely in temporal cues (obtained by manipulating the phase relationship between components) can elicit streaming. Finally, Grimault *et al.* (2002) observed streaming based on the modulation rate of sinusoidally amplitude-modulated noises, i.e., without any spectral cues to pitch. Despite the fact that the pitch elicited by the modulated noise was relatively weak, these authors observed streaming similar to that obtained with unresolved complex tones. Thus, streaming is reduced when spectral cues are reduced, but it is apparently possible to some extent when spectral cues are removed and only temporal cues remain.

These results have substantial implications for individuals with sensorineural hearing impairment and those fitted with a CI. It is well known that these individuals have reduced access to spectral cues (cf. Moore, 1998). Fundamental frequency DLs are approximately 2.5 times greater than normal in HI listeners (Moore and Peters, 1992) and 7.7 times greater in CI users (Rogers *et al.*, 2006). These results suggest that pitch differences are far less salient for these listeners, and that pitch-based streaming might be impaired. Indeed a few studies argue that reduced frequency selectivity is responsible for the relatively poor performance of CI users in the perception of concurrent voices (Qin and Oxenham, 2005; Stickney *et al.*, 2004, 2007). However, psychoacoustic measures have indicated that temporal resolution is generally intact in the HI ear (for review, see Moore, 1998; Healy and Bacon, 2002). CI users are also sensitive to temporal rate pitch, up to a limit of approximately 300 Hz (Shannon, 1983; Tong and Clark, 1985; Townshend *et al.*, 1987). Although their DLs for rate discrimination are larger than in NH (Zeng, 2002) CI listeners can use this cue to discriminate vowel F_0's (Geurts and Wouters, 2001). Although these results indicate that the cues for streaming may be available, their use by these individuals is not well understood.

Auditory streaming has been examined to a limited extent in HI listeners, with mixed results. Grose and Hall (1996) found that listeners with cochlear hearing loss generally required a greater frequency separation for segregation of pure tones. However, Rose and Moore (1997) reported no systematic difference between ears of unilaterally impaired listeners in this task. The correlation between auditory filter width and pure-tone streaming was also found to be not significant (Mackersie *et al.*, 2001). Grimault *et al.* (2001) found that streaming was hindered for HI listeners relative to NH, but only in conditions where components of complex tones were resolved for NH and unresolved for HI listeners. Finally, Stainsby *et al.* (2004) examined streaming based on phase relationship differences and found results for elderly HI listeners that were similar to those observed in NH listeners.

A few studies have also attempted to examine streaming in CI users (Hong and Turner, 2006; Chatterjee *et al.*, 2006; Cooper and Roberts, 2007). For these users, different kinds of temporal cues can be related to pitch. Moore and Carlyon (2005) argued that the temporal fine structure of resolved harmonics was the most accurate pitch mechanism. However, when harmonics are unresolved, they interact in auditory filters and can encode pitch by amplitude modulation (AM) rate (i.e., by the temporal envelope periodicity). Because of the way the spectrum is partitioned in the CI processor, harmonics of lower pitched human voices ($F_0 \sim 100$ Hz) almost always interact in the first channel of the CI processor. Thus, the availability of individual resolved harmonics is extremely limited. In contrast, the temporal envelope is roughly preserved in each band, so pitch may be coded by temporal periodicity cues. In this paper, the term "temporal-pitch cues" will then refer to temporal periodicity (in the range 100–400 Hz); in contrast to "spectral-pitch cues," which will refer to the pitch that is evoked by resolved harmonics (i.e., issued from the tonotopic analysis of the cochlea, and if relevant, from some analysis of temporal fine structure). Because amplitude-modulated broadband noises can produce some impression of pitch (Burns and Viemeister, 1976, 1981) and can induce streaming (Grimault *et al.*, 2002), it might be possible for these temporal cues to induce streaming in CI users.

Hong and Turner (2006) used the rhythm task described in Roberts *et al.* (2002) to obtain an objective measure of obligatory streaming in NH and CI users. They found that half of the 16–22 electrode CI users performed as poorly as the NH listeners (suggesting streaming), whereas the other half performed better than normal (suggesting less stream segregation). The authors showed that this variability correlated moderately but significantly with the ability to perceive speech in noise. Chatterjee *et al.* (2006) used pulse trains in *ABA* patterns and a subjective evaluation of whether subjects fitted with the 22-channel nucleus CI heard one or two streams. These authors observed response patterns that could be explained by streaming for both differences in spatial location (presentation electrode) and AM rate (in a single subject). However, they did not observe the characteristic buildup of streaming over time (Bregman, 1978) for simple pulsatile stimuli that differed in location. This observation raises the possibility that the task involved discrimination rather than streaming. On the contrary, they did observe some buildup for the signals that differed in AM rate, which suggests that AM rate based streaming was indeed observed. Cooper and Roberts (2007) also employed pulsatile stimuli that differed in electrode location. They obtained subjective reports involving the presence of two streams, but a second experiment revealed that the results may have been attributable to pitch (or brightness) discrimination. Other studies have targeted temporal cues more specifically, but have examined simultaneous segregation by CI users. Using a CI simulation based on filtered harmonic complexes, Deeks and Carlyon (2004) found only modest improvements in concurrent sentence recognition when the target and masker were presented at different pulse rates. Also, Carlyon *et al.* (2007) found that a difference in rate pitch did not enhance simultaneous segregation of pulse trains in CI users. Altogether, these studies provide only modest evidence that segregation or streaming can occur in CI recipients on the basis of either place pitch (i.e., electrode number) or temporal pitch.

These previous results together suggest (1) that F_0-based streaming is affected by frequency selectivity, but (2) that

streaming can be also induced by temporal-pitch cues. It is also clear that (3) frequency selectivity is reduced in HI and CI listeners, but that (4) temporal-pitch cues are preserved to some extent in these listeners. The question then becomes to what extent these cues can be utilized to elicit streaming.

Although streaming is often assumed to be a primitive mechanism, some correlation between streaming and higher level processing, such as concurrent speech segregation, has been reported (Mackersie *et al.*, 2001). However, the relation between streaming with pure or complex tones and speech segregation remains difficult to assess. In speech, pitch cues signaling that different talkers are present are mixed with other cues that may not be relevant for concurrent talker segregation. Listeners may then not benefit from these cues in ecological situations. Only a few studies have reported streaming with speech materials (Dorman *et al.*, 1975; Nooteboom *et al.*, 1978; Tsuzaki *et al.*, 2007; Gaudrain *et al.*, 2007), and only the last one examined the effect of impaired frequency selectivity.

The current study follows that of Gaudrain *et al.* (2007). Whereas streaming under conditions of broad tuning in accord with sensorineural hearing impairment was examined in that study, streaming under conditions similar to CI stimulation was assessed in the current study. Specifically, the role of reduced spectral-pitch cues in streaming of speech stimuli was assessed in a first experiment, and the possible role of temporal-pitch cues was investigated in a second experiment.

In this study, noise-band vocoder models of CIs were employed to control the spectral and temporal cues available to listeners. The use of NH listeners exposed to cues reduced in controlled manner allowed the elimination of complications and confounds associated with the clinical population. In addition, an objective paradigm—the order task—was used to assess streaming. In this task, the listener is presented a repeating sequence of vowels having alternating F_0 and asked to report the order of appearance of the constituent vowels (Dorman *et al.*, 1975). If the sequence splits into streams corresponding to the two F_0's, the loss of temporal coherence across streams hinders the ability to identify the order of items within the sequence. Although the order of items within individual steams is available to the listener, the order of items across streams is not. Thus, this task requires that the subject *resists* segregation to perform well. As a result, the task is used to assess "obligatory" streaming—that which cannot be suppressed by the listener (see Gaudrain *et al.*, 2007).

This type of streaming does not produce a substantial cognitive load, and it is less dependent on attention and subject strategy than the subjective evaluation of one versus two streams. In addition, this approach is appropriate for examining segregation in the presence of reduced cues—because performance tends to improve (less streaming) with degraded stimuli, performance cannot be attributable to the degradation of the individual items. The reduction in spectral resolution associated with the vocoder is expected to reduce the amount of streaming. Consequently, performance in the order task should improve in the CI simulation, relative to the intact stimuli. However, if temporal-pitch cues encoded by the CI simulation are sufficient to elicit obligatory streaming, overall scores should remain low and an effect of F_0 separation should be observed.

II. EXPERIMENT 1

A. Materials and method

1. Subjects

Six subjects aged 22–30 years (mean 26.2) participated. All were native speakers of French and had pure-tone audiometric thresholds below 20 dB hearing level (HL) at octave frequencies between 250 and 4000 Hz (American National Standards Institute, 2004). All were paid an hourly wage for participation. These subjects participated in one of the experiments of Gaudrain *et al.* (2007) and were therefore familiar with the paradigm.

2. Stimuli

Individual vowels were first recorded and processed, then arranged into sequences. The six French vowels /a e i ɔ y u/ were recorded (24 bits, 48 kHz) using a Røde NT1 microphone, a Behringer Ultragain preamplifier, a Digigram VxPocket 440 soundcard, and a PC. The speaker was instructed to pronounce all six vowels at the same pitch and to reduce prosodic variations. The F_0 and duration of each vowel were then manipulated using STRAIGHT (Kawahara *et al.*, 1999). Duration was set to 167 ms to produce a speech rate of 6.0 vowel/s. This value is close to that measured by Patel *et al.* (2006) for syllable rates in British English (5.8 syllables/s) and in French (6.1 syllables/s). Additional versions of each vowel were then prepared in which the average F_0's were 100, 110, 132, 162, and 240 Hz. Fundamental frequency variations related to intonation were constrained to be within 0.7 semitones (4%) of the average. This value was chosen to allow F_0 variations within each vowel, but to avoid overlap across the F_0 conditions. Formant positions were held constant across F_0 conditions.

Each vowel was subjected to two conditions of reduced spectral resolution. In Q_{20} the vowels were subjected to a 20-band noise vocoder, and in Q_{12} they were subjected to a 12-band noise vocoder. The Q_{12} condition was intended to be closer to actual CI characteristics, while the Q_{20} condition was intended to be an intermediate condition with more spectral detail. Q_∞ refers to the intact vowels. The implementation of the noise-band vocoder followed Dorman *et al.* (1997). The stimulus was first divided into frequency bands using eighth order Butterworth filters. The cutoff frequencies of these bands were the approximately logarithmic values used by Dorman *et al.* (1998) and are listed in Table I. The envelope of each band was extracted using half-wave rectification and eighth order Butterworth lowpass filtering with cutoff frequency of 400 Hz. This lowpass value ensured that temporal-pitch cues associated with voicing were preserved. The resulting envelopes were used to modulate white noises using sample point-by-point multiplication, which were then filtered to restrict them to the spectral band of origin. The 12 or 20 bands comprising a condition were then mixed to construct the vocoder. A 10 ms cosine rise/fall gate was finally applied to each vowel in each condition.

TABLE I. Cutoff frequencies (Hz) of the 12- and 20-channel vocoders, from Dorman *et al.* (1998).

Channel	Q_{20}	Q_{12}
1	166	212
2	198	336
3	362	570
4	414	754
5	600	1056
6	676	1324
7	888	1718
8	992	2098
9	1236	2620
10	1376	3150
11	1658	3848
12	1840	4582
13	2168	
14	2400	
15	2786	
16	3078	
17	3534	
18	3900	
19	4438	
20	4894	
	5532	

The vowels were then concatenated to form sequences. Figure 1 describes the arrangement of vowels into sequences and the construction of the various conditions. Each sequence contained one presentation of each vowel. Sequences containing all possible arrangements of the six repeating vowels ($[n-1]!=120$) were first generated, then the 60 arrangements having the smallest differences in formant structure were selected for inclusion [Fig. 1(A)]. The selection of arrangements having the smallest *perceptual formant differences*[1] was performed to reduce the influence of streaming based on differences in formant structure between successive vowels in a sequence (Gaudrain *et al.*, 2007). These 60 arrangements were then divided into five groups of 12 arrangements each, such that the average perceptual distance of each group was approximately equal [Fig. 1(B)].

The F_0 of the vowels in a sequence alternated between two values $F_{0(1)}$ and $F_{0(2)}$. In condition LowRef, the value of $F_{0(1)}$ was 100 Hz and $F_{0(2)}$ was one of the five F_0 values (100, 110, 132, 162, and 240). In condition HiRef, $F_{0(1)}$ was 240 Hz and $F_{0(2)}$ was one of the five F_0 values. Thus, there were five F_0 differences. Each group of 12 arrangements was then assigned to one of the five F_0 differences. The appearance of the same 60 arrangements in both the LowRef and HiRef conditions yielded 120 sequences [Fig. 1(C)]. These 120 sequences appeared in both a Slow and a Fast condition, yielding 240 sequences [Fig. 1(D)]. Finally, each of these 240 sequences appeared in each of the three Q conditions, yielding a total of 720 sequences [Fig. 1(E)].

In the Slow condition, the presentation rate was 1.2 vowel/s, and in the Fast condition, it was 6 vowel/s. Slow sequences were used to check vowel identification performance, and Fast sequences were used to examine streaming.

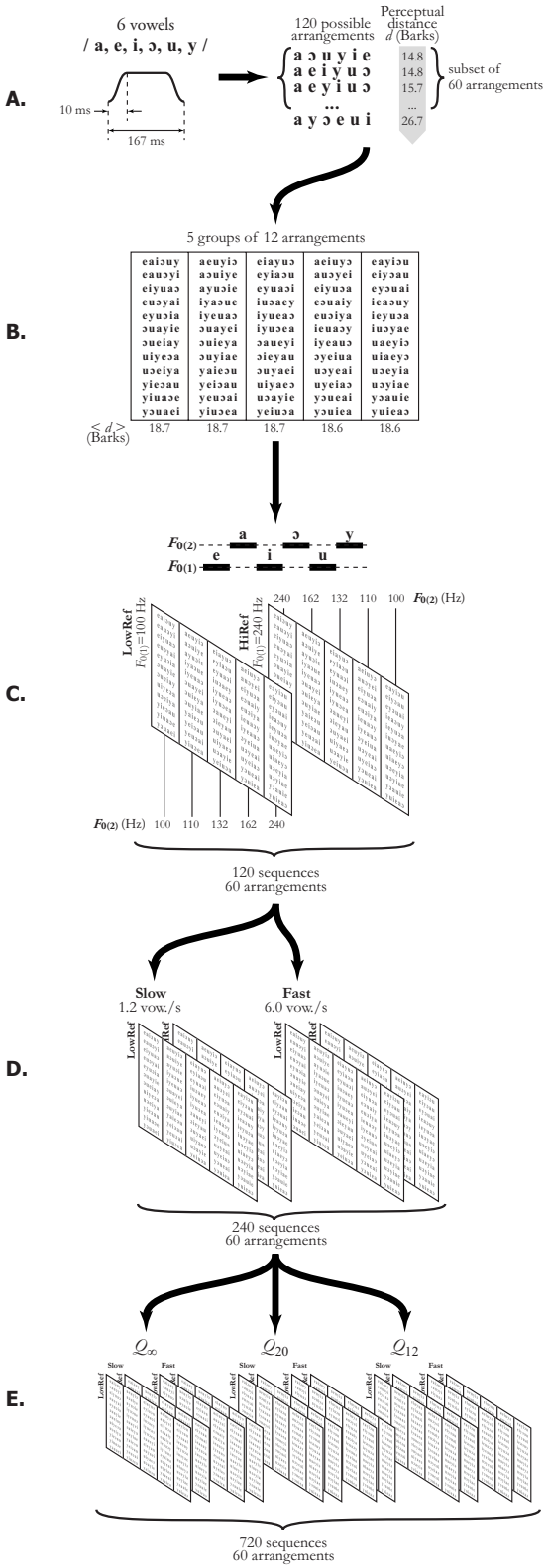

FIG. 1. The arrangement of conditions: (A) Individual vowels were recorded and modified using STRAIGHT. They were arranged into sequences (120 possible orders), and the 60 arrangements having the lowest perceptual distances d (in barks) were selected (see text for details). (B) These 60 arrangements were divided into five groups with similar average perceptual distance (12 in each group). (C) Each group was assigned to a fundamental frequency difference in both LowRef and HiRef conditions (yielding 120 sequences). (D) These 120 sequences appeared in both Slow and Fast conditions (yielding 240 sequences). (E) Finally, these 240 sequences appeared in each Q condition (yielding 720 sequences). These 720 sequences were presented across six presentation blocks, such that each condition was equally represented within each block.

To create the Slow sequences, silence was added between the vowels so that vowel duration remained constant across rate conditions. The Slow sequences were repeated four times and the Fast sequences were repeated 20 times, for overall stimulus durations of 20 s.

Stimuli were generated at 16 bits and 44.1 kHz using MATLAB. They were presented using the Digigram VxPocket 440 soundcard, and Sennheiser HD250 Linear II headphones diotically at 85 dB sound pressure level, as measured in an artificial ear (Larson Davis AEC101 and 824; American National Standards Institute, 1995).

3. Procedure

a. Training and selection. Two training tasks preceded testing. The first involved simple identification of single vowels. Subjects heard blocks containing each vowel at each F_0 twice. They responded using a mouse and a computer screen, and visual feedback was provided after each response. This test was repeated, separately for each Q condition, until a score of 98% (59/60) was obtained. On average, proficiency was reached after one block for the Q_∞ vowels, 1.3 blocks for the Q_{20} vowels, and 2.8 blocks for the Q_{12} vowels.

The second training task involved vowel identification using the Slow sequences. In each block, 60 sequences were presented representing all 30 conditions (5 F_0's $\times 2$ LowRef/HiRef $\times 3$ Q conditions). The procedure was the same as the test procedure, except that visual feedback was provided. The subject was presented with a repeating sequence. After an initial 5 s period, during which streaming was allowed to stabilize, the subject was asked to report the order of appearance of the constituent vowels. They were allowed to start with any vowel. The response was entered using a computer graphic interface and a mouse. The next sequence was presented after the subject confirmed their response or after a maximum of 20 s. Visual feedback was then provided. To proceed to the test, subjects were required to obtain a score, averaged over two consecutive blocks, greater than 95% in each Q condition. On average, 6.3 blocks were necessary to reach the proficiency criterion. Although intended to be a selection criterion, no subject was eliminated at this step.

b. Streaming test. The procedure was the same as that in the second training task, except that no feedback was provided. The 720 sequences were distributed among six presentation blocks, such that each condition was represented equally in each block. The average duration of one block was approximately 28 min. Experiment 1 required subjects to participate in four 2 h sessions, during which frequent breaks were provided. The experimental procedure was formally approved by a local ethics committee (CCPPRB Léon Bérard).

B. Results

For each condition, the score is the percentage of responses in which the six vowels comprising a sequence were reported in the correct order. Mean scores across subjects are plotted as a function of $F_{0(2)}$ in Fig. 2. Chance performance is 0.8%. As in Gaudrain *et al.*, 2007, high scores can be interpreted as a tendency toward integration across $F_{0(1)}$ and $F_{0(2)}$ items and a resistance to obligatory streaming. Separate analyses were conducted on the LowRef and HiRef conditions because the F_0 differences were not the same in the two conditions. The results in the Slow condition (1.2 vowel/s) showed that identification was near perfect in all conditions except one (HiRef, Q_{12}, $F_{0(2)}=162$ Hz). An analysis of errors in this condition showed confusions between /y/ and /e/ in 8/9 false responses. These two vowels therefore seem difficult to discriminate at this particular combination of F_0's and vocoder channel divisions. All subsequent analyses were carried out on the data collected in the Fast conditions.

A two-way analysis of variance (ANOVA) on the Fast/LowRef data using Q condition and $F_{0(2)}$ as repeated parameters indicated that the effects of Q condition [$F(2,10)=4.14$, $p<0.05$] and $F_{0(2)}$ [$F(4,20)=12.21$, $p<0.001$] were significant, and interacted significantly [$F(8,40)=3.93$, $p<0.01$]. Separate one-way ANOVAs on each Q condition using $F_{0(2)}$ as a repeated factor showed a significant effect of $F_{0(2)}$ in the Q_∞ condition [$F(4,20)=9.28$, $p<0.001$], but not in the Q_{12} [$F(4,20)=0.65$, $p=0.63$] or Q_{20} conditions [$F(4,20)=0.21$, $p=0.93$].

A two-way ANOVA on the Fast/HiRef data using Q condition and $F_{0(2)}$ as repeated parameters indicated that Q condition [$F(2,10)=9.02$, $p<0.01$] and $F_{0(2)}$ [$F(4,20)=14.30$, $p<0.001$] were significant, and interacted significantly [$F(8,40)=6.74$, $p<0.001$]. Separate one-way ANOVAs on each Q condition using $F_{0(2)}$ as a repeated factor showed significant effects of $F_{0(2)}$ in the Q_∞ condition [$F(4,20)=10.15$, $p<0.001$] and in the Q_{12} condition [$F(4,20)=14.96$, $p<0.001$], but not in the Q_{20} condition [$F(4,20)=1.50$, $p=0.24$]. A *post hoc* analysis using pairwise t-tests showed that the effect of $F_{0(2)}$ in the Q_{12} condition was due solely to the point $F_{0(2)}=162$ Hz. As previously stated, the confusions in this particular condition suggest difficulty with this particular set of parameters. When the HiRef condition was analyzed with this condition excluded, the pattern of significance was identical to that observed in the LowRef conditions: a significant effect of $F_{0(2)}$ in the Q_∞ condition [$F(3,15)=12.24$, $p<0.001$], but not in the Q_{12} [$F(3,15)=0.74$, $p=0.54$] or Q_{20} conditions [$F(3,15)=1.13$, $p=0.37$].

C. Discussion

The results in the natural speech condition (Q_∞) are consistent with those observed by Gaudrain *et al.* (2007) in their first experiment. The greater the F_0 difference, the lower the scores, signifying greater streaming. Streaming based on F_0 difference is considered to be *obligatory* here because the task employed required that streaming be suppressed in order to perform accurately. Although the pattern of results in the current experiment is similar to that obtained by Gaudrain *et al.* (2007), the baseline level of performance differs. Scores in the Q_∞ condition at matched F_0 were over 80% here and approximately 50% in experiment 1 of Gaudrain *et al.* (2007). One possible reason is that participants in the current experiment were well trained. In addition, there were subtle differences in the stimuli used in the two studies.

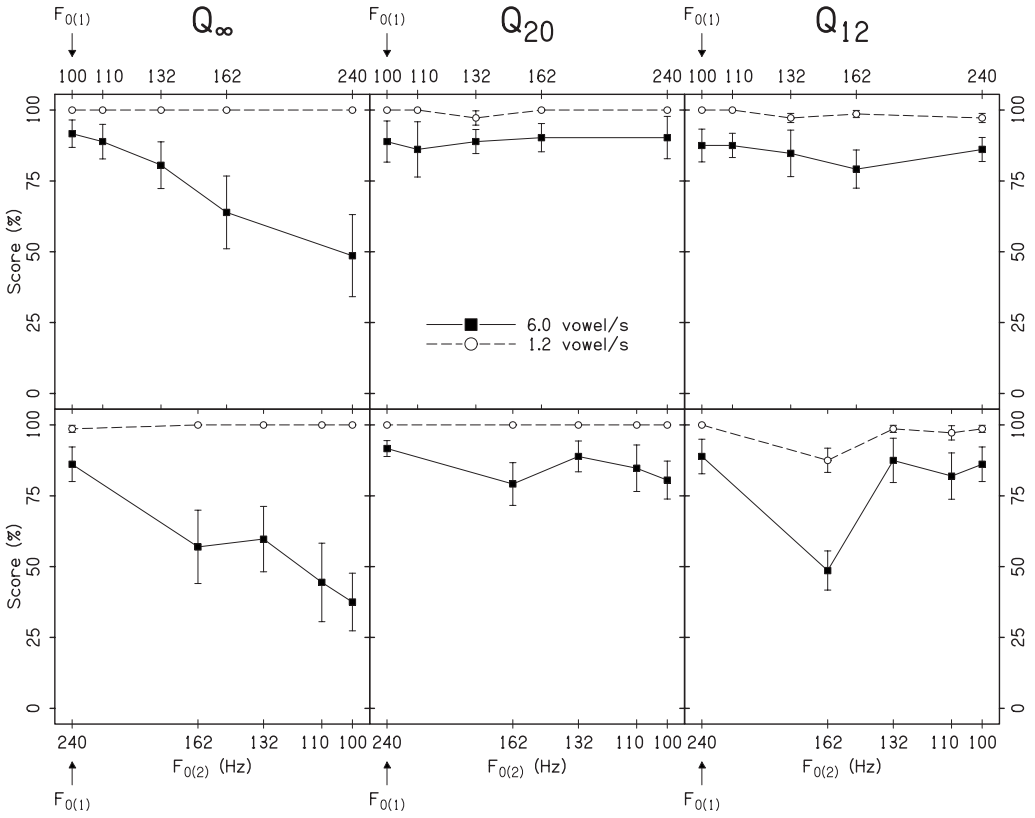

FIG. 2. Shown are group means and standard deviations for a task in which subjects reported the order of six vowels appearing in sequence. Thus, low scores represent a tendency toward segregation. Alternate vowels were at alternate F_0 values ($F_{0(1)}$ and $F_{0(2)}$). In the LowRef conditions (upper panels), $F_{0(1)}$ was fixed at 100 Hz and in the HiRef conditions (lower panels) $F_{0(1)}$ was fixed at 240 Hz. Conditions Q_{20} and Q_{12} involved reduced frequency resolution via 20- and 12-channel noise vocoders. Filled squares represent a Fast condition (6.0 vowel/s) in which streaming can occur, and open circles represent a Slow condition (1.2 vowel/s) that ensures accurate item identification. The abscissa is logarithmic.

Gaudrain *et al.* (2007) attributed low scores in the matched F_0 condition to formant-based streaming. Such a phenomenon has been reported by Dorman *et al.* (1975) with synthesized vowels. Formant-based streaming might be reduced with the recorded vowels used in the current experiment, where small F_0 fluctuations were preserved. Fundamental frequency fluctuations might serve to strengthen the grouping of components comprising individual vowels and limit the grouping of formants across successive vowels, as suggested by Gestalt theory (Bregman, 1990).

In the conditions having spectral degradation (Q_{20} and Q_{12}), the scores are high and do not depend on F_0. Thus, when spectral cues to pitch were reduced in accord with a CI model, F_0-based streaming was reduced or eliminated. Further, these results indicate that the temporal cues to pitch that remained in the vocoded stimuli were not strong enough, in this case, to elicit obligatory streaming. It is potentially interesting to note that these results cannot be explained by a loss of intelligibility since degradation of the stimuli yielded an increase in performance. In addition, vowel identification was confirmed in the Slow condition.

The main finding of this experiment was that no F_0-based streaming appeared when spectral-pitch cues were degraded using a model of a 12- or a 20-channel CI. This result suggests that obligatory streaming is reduced when spectral cues to pitch are reduced in this manner and may not be possible for vowel stimuli based on remaining temporal-pitch cues. This observation is in apparent contrast with studies that observed some streaming in CI recipients (Chatterjee *et al.*, 2006; Hong and Turner, 2006; Cooper and Roberts, 2007). However, these previous observations were generally based on conditions in which some place pitch existed. The current result is also in apparent contrast with the observation of streaming based on temporal pitch in NH (Grimault *et al.*, 2002). One explanation for this discrepancy is that temporal-pitch cues were not sufficiently salient in the noise-band vocoder. This point is addressed in Sec. IV. It is also potentially important that obligatory streaming is strongly influenced by presentation rate (van Noorden, 1975), and that Hong and Turner (2006), Chatterjee *et al.* (2006), and Grimault *et al.* (2002) all used sequences with higher presentation rates (10 stimuli/s) to observe streaming. It then seems plausible that, for an F_0 difference of about one octave, the natural presentation rate used in experiment 1 was not sufficiently high to elicit obligatory streaming under degraded conditions, but that streaming may still be possible. The next experiment assesses this hypothesis.

III. EXPERIMENT 2

As shown by van Noorden (1975), the temporal coherence boundary, the threshold corresponding to obligatory streaming, depends on presentation rate. As shown in experiment 1 of Gaudrain *et al.* (2007), higher presentation rates in

the current paradigm do indeed lead to stronger measures of streaming. Thus, increasing the repetition rate should strengthen the streaming effect and reveal if segregation is possible under the current conditions of severely reduced spectral cues, but preserved temporal cues to pitch. In addition, two envelope cutoff values were employed to more closely examine the role of temporal cues.

A. Materials and method

1. Subjects

Nine subjects aged 18–27 years (mean 21.9) participated. All were native speakers of French and had NH as defined in experiment 1. None of these subjects participated in previous similar experiments, and all were paid an hourly wage for participation.

2. Stimuli

The same six recorded vowels employed in experiment 1 were used. The durations of the vowels were reduced to 133 ms using STRAIGHT. Again, 10 ms ramps were employed. The average F_0's of each vowel were set to 100, 155, and 240 Hz using the same method used in experiment 1. The intact vowels were used for a Q_∞ condition. The same noise-band vocoder used in experiment 1 was again used to process vowels for Q_{20} and Q_{12} conditions. However, unlike experiment 1, two cutoff frequencies (f_c) were used for envelope extraction. A value of f_c=400 Hz was employed to preserve temporal-pitch cues, and a value of f_c=50 Hz was employed to eliminate temporal-pitch cues. As in experiment 1, envelope extraction involved half-wave rectification and eighth order Butterworth lowpass filtering.

The processed vowels were then concatenated to form sequences in the five processing conditions (Q_∞, Q_{20} f_c =400 Hz, Q_{20} f_c=50 Hz, Q_{12} f_c=400 Hz, and Q_{12} f_c =50 Hz). The 36 arrangements having the lowest perceptual distance were selected and divided into three groups having approximately equal mean perceptual distance values. As in experiment 1, these three groups were used for the three F_0 separation conditions. Thus, as in experiment 1, the particular arrangements of vowels were distributed across the F_0 separation conditions, but repeated across the other conditions to ensure that effects associated with the particular order of items were constant.

In this experiment, only the LowRef condition was used, so that $F_{0(1)}$ was always 100 Hz. The low identification score observed for $F_{0(2)}$=162 Hz in the HiRef condition of experiment 1 was hence avoided. As in the first experiment, Slow (1.2 vowel/s) and Fast (7.5 vowel/s) sequences were employed. Slow sequences were repeated 4 times and Fast sequences were repeated 25 times, so that the overall duration in both conditions was 20 s. Stimuli were generated with MATLAB as 16 bit, 44.1 kHz sound files, and were presented as in experiment 1.

3. Procedure

a. Training and selection. Training again began with simple identification of single vowels. Five blocks of 72 vowels were presented. Each block contained four repetitions of each vowel at each $F_{0(2)}$ in a single degradation condition, in random order (4 repetitions $\times$ 6 vowels $\times$ 3 F_0's=72 items). The blocks were presented from the least degraded (Q_∞) to the most degraded (Q_{12} f_c=50 Hz). Visual feedback was provided. Each block was repeated until a score of 96% (69/72) was obtained or a maximum of three repetitions was reached. On average the blocks were repeated 2.0 times for each condition (range 1.7–2.3).

As in experiment 1, training involving the Slow condition sequences followed. The test consisted of seven blocks. Each block was composed of 36 sequences and all the conditions were represented at least twice in random order. Subjects were required to score greater than 95% correct over three successive blocks in each condition to advance to the next stage. One subject was unable to reach the criterion and was dismissed. For seven of the remaining participants, five blocks were sufficient to reach the criterion. The last subject reached the criterion after seven blocks.

b. Streaming test. The test consisted of 5 blocks of 72 sequences each. All conditions (3 $F_{0(2)}$'s, 5 Q's, and 2 Slow/Fast) were represented as equally as possible in each block. For each $F_{0(2)}$, streaming was measured over 12 different arrangements of vowels. Other aspects of the experiment, including the initial 5 s response lockout and the manner of response, were identical to those of experiment 1. Experiment 2 required subjects to participate in three to four 2 h sessions, during which frequent breaks were provided. The experimental procedure was formally approved by a local ethics committee (CPP Sud Est II).

B. Results

Results averaged across subjects are plotted in Fig. 3. As can be seen, scores were uniformly high in the Slow conditions (mean: 98.6% correct), reflecting accurate identification of the constituent items. The subsequent analyses were conducted on the Fast conditions. A two-way ANOVA, using processing condition (Q_∞, Q_{20} f_c=400 Hz, Q_{20} f_c=50 Hz, Q_{12} f_c=400 Hz, and Q_{12} f_c=50 Hz) and F_0 as repeated parameters, showed a significant effect of processing condition [$F(4,28)=23.57$, $p<0.001$] as well as F_0 [$F(2,14)=8.25$, $p<0.01$]. These factors did not interact [$F(8,56)=1.31$, $p=0.26$], indicating that the effect of F_0 was not significantly different between processing conditions. To test for the effect of f_c, separate two-way ANOVAs (using F_0 and f_c as repeated parameters) were performed in each of the degraded conditions. The analyses revealed no effect of f_c in the Q_{12} [$F(1,7)=0.9$, $p=0.37$] or Q_{20} conditions [$F(1,7)=3.86$, $p=0.09$]. The effect of F_0 was significant in the Q_{12} condition [$F(2,14)=7.11$, $p<0.01$], but not in the Q_{20} condition [$F(2,14)=0.47$, $p=0.63$]. The interaction was not significant in the Q_{12} condition [$F(2,14)=0.78$, $p=0.48$], but was at Q_{20} [$F(2,14)=6.37$, $p<0.01$], as suggested by the pattern of results in Fig. 3. An analysis of this interaction using pairwise t-tests revealed no significant difference between f_c values, even at $F_{0(2)}$=100 Hz [$p=0.30$].

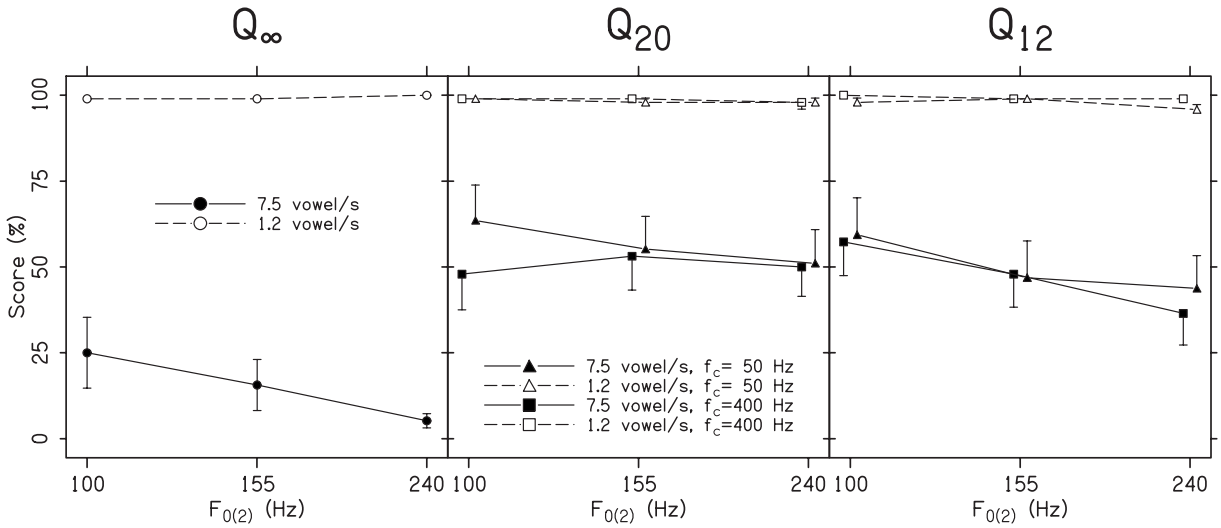

FIG. 3. Shown are group means and standard deviations for the vowel order identification task in experiment 2. Again, low scores represent a tendency toward segregation. $F_{0(1)}$ was at 100 Hz and $F_{0(2)}$ is shown. In each Q condition, scores for the Slow conditions (1.2 vowel/s) are plotted as open symbols, and scores in the Fast conditions (7.5 vowel/s) are plotted as filled symbols. In the conditions Q_{20} and Q_{12} (middle and rightmost panels), scores for conditions in which temporal smoothing (f_c) was 50 Hz are plotted with triangles, and scores for f_c=400 Hz are plotted with squares. The abscissa is logarithmic.

C. Discussion

The scores in the matched F_0 conditions were lower in the current experiment than in experiment 1, indicating that the subjects performed more poorly in this second experiment. This result can likely be attributed to at least two sources. The first is the use of naive listeners in the current experiment and trained listeners in experiment 1. The second is that increasing the presentation rate enhances segregation based on formant structure. Gaudrain *et al.* (2007) argued that vowels having matched F_0 can elicit streaming based on formant structure, as found by Dorman *et al.* (1975). The higher scores in the Q_{20} and Q_{12} conditions support this hypothesis, suggesting, as in Gaudrain *et al.*, 2007, that formant-based streaming is hindered by loss of frequency resolution. Despite the fact that scores were reduced in the $F_{0(2)}$=100 Hz conditions, a significant main effect of $F_{0(2)}$ was nevertheless observed across all conditions. Although this effect was not observed at Q_{20}, it was observed in the isolated Q_{12} condition, indicating that some F_0-based streaming occurred in this degraded condition.

The presence (f_c=400 Hz) or absence (f_c=50 Hz) of temporal-pitch cues did not influence the performance in either degraded condition. The use of an eighth order smoothing filter during envelope extraction ensured that modulation frequencies above these temporal cutoffs were not present at meaningful levels (Healy and Steinbach, 2007). This result suggests that streaming was not based on temporal-pitch cues. Thus, other reasons for F_0-based streaming must be sought.

One obvious F_0-related cue that can be considered is the modulation product. Because of basilar membrane nonlinearity, the periodic modulation in the channels of the vocoder could evoke a component at this modulation frequency and at harmonic multiples. If the envelope modulation periodicity represents the F_0, the modulation product can recreate the original first harmonic. However, when f_c=50 Hz, pitch cues associated with F_0 are removed from the envelope. If amplitude modulation had elicited streaming, a difference should have been observed between the f_c conditions. Thus, although modulation products may be evoked with the present stimuli, they are likely not responsible for the observed pitch-based streaming.

It must then be considered that some remaining spectral cues could be related to the F_0. One such cue involves the first harmonic. The first harmonic can fall either in the first band of the vocoder or below it. It could then possibly provide an F_0-related cue because it can influence the level of the first channel. To test this hypothesis, the level in the first channel was measured in the different conditions. A three-way ANOVA on this measure using F_0, Q, and f_c as repeated parameters, across the six vowels, showed no effect of F_0 [$F(2,10)=0.03$, $p=0.97$] (f_c [$F(1,5)=0.84$, $p=0.40$] and Q [$F(1,5)=449.3$, $p<0.001$]). Thus, the first channel does not appear to contain a consistent F_0-related cue. However, the distribution of harmonics across all channels could have an effect. A Fisher discriminant analysis (FDA) was used to find a linear combination of channel levels that provides the best representation of the F_0 in the Q_{12} condition (using Python MDP, Berkes and Zito, 2007). As shown in Fig. 4, for both f_c=400 and 50 Hz, a linear combination of channel levels can be found to represent the F_0. This linear combination can be used as a metric that represents F_0. In Fig. 4, scores were plotted against this metric to show the relation between this metric and segregation. In conclusion, it is possible that some spectral cues related to the F_0, but not capable of generating a strong pitch sensation, could persist even in the absence of harmonics.

Although the evidence for remaining F_0-related spectral cues in the Q_{12} condition is relatively clear, it remains unclear whether these cues would be present in a real CI and for sounds other than those employed here. It is then important to determine the origin of these cues. The vocoder discards the harmonic structure of the vowels. Hence, if a spectral cue is preserved by the vocoder, it must be encoded by

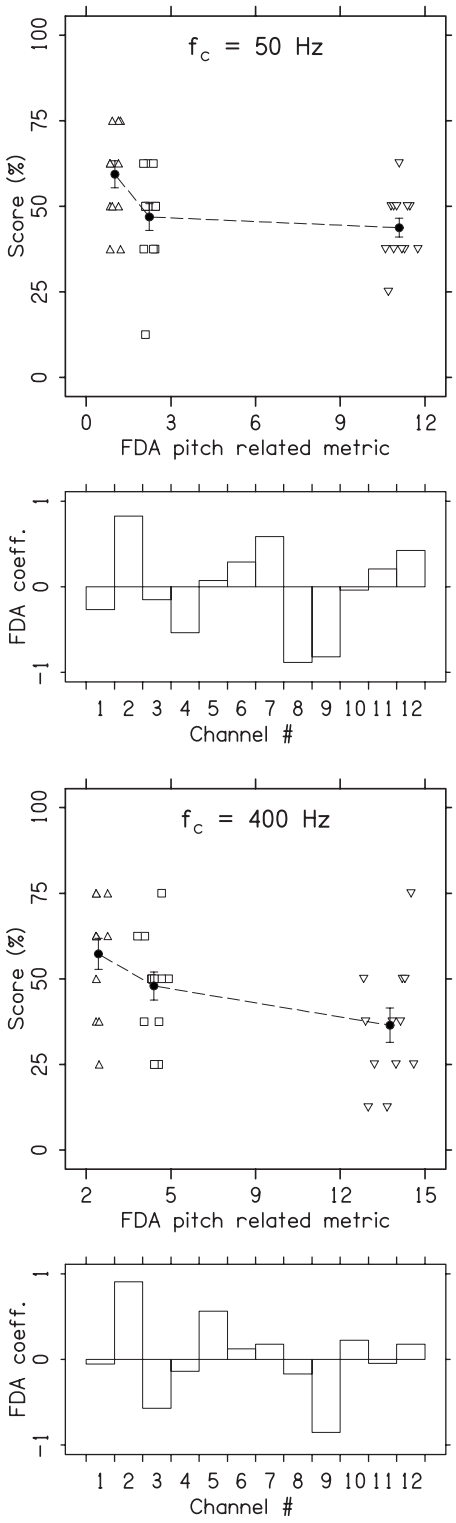

FIG. 4. Upper panels: Scores as a function of the pitch related metric found by FDA on vocoder channel mean levels for vowels processed in the Q_{12} condition, f_c=50 Hz. The upward triangles represent sequences with $F_{0(2)}$ =100 Hz, squares for $F_{0(2)}$=155 Hz, and downward triangles for $F_{0(2)}$ =240 Hz. The coefficients of the linear combination used as a metric are represented in the lower subpanel. Lower panels: Same representation for the Q_{12} condition, f_c=400 Hz.

the spectral envelope that is partially preserved. To examine this, the spectral envelopes of the unprocessed vowels (Q_∞) were extracted using STRAIGHT for the three different F_0's. These envelopes showed slight but consistent differences in the formant regions that are probably attributable to the formant definition in relation to the harmonic density: Higher F_0's tended to produce broader formants. This small broadening effect seems to have been emphasized by the spectral quantization in the 12-channel vocoder.

It would have been interesting to determine whether this cue existed in the Q_{20} condition. However, it was not possible to analyze the Q_{20} condition because FDA requires the number of items (6 vowels $\times$ 3 F_0's=18) to be greater than the number of parameters (20 vocoder bands). It is possible that the greater number of channels reduced the quantization effect, and did not emphasize this F_0-related cue as much as in the 12-channel vocoder.

This analysis of the stimuli suggests that the F_0-related spectral cues are due to the modification of formant definition associated with changes in F_0. Although this phenomenon may occur when perceiving natural vowels uttered at different F_0's, the frequency quantization of the 12-channel noise-band vocoder simply emphasized the effect, as well as making it relatively stochastic across vowels. It is worth noting that, in the current experiment, the formant positions were held constant while changing F_0. In real speech, the formant position and width are related to speaker size, which also drives the nominal F_0 of the speaker. This F_0-related cue may then appear along with some other cues associated with vocal tract length (VTL), which can work to further support streaming. Although streaming in NH listeners seems to be less influenced by changes in VTL than in F_0 (Tsuzaki *et al.*, 2007), this kind of cue could be preserved and used in CI listening.

IV. GENERAL DISCUSSION

The main result of this study is that obligatory streaming was reduced, but still present, with spectral cues reduced in accord with a CI model, but that this streaming was not attributable to temporal-pitch cues. This is consistent with Gaudrain *et al.* (2007) who observed that impoverishing the spectral representation of vowels via simulated broadened auditory tuning hindered pitch-based streaming. Reducing the spectral cues and preserving most of the temporal cues in experiment 1 reduced sufficiently the salience of the two different streams to eliminate obligatory streaming. In the second experiment, emphasizing streaming through a higher presentation rate did not lead to the observation of streaming based on temporal cues, but instead led to the observation of streaming based on a stochastic spectral cue.

The current results are also consistent with previous observations of concurrent speech perception in NH listeners hearing CI simulations. NH listeners are able to take advantage of a pitch difference between a target and a masker to enhance speech recognition in noise (Brokx and Nooteboom, 1982). However, NH listeners are less able to take advantage of pitch differences between speech target and masker when the stimuli are processed with a noise-band vocoder (Stickney *et al.*, 2004). Similarly, Deeks and Carlyon (2004) reported that pulse rate differences did not enhance the reception of concurrent speech in a CI simulation based on unresolved complexes. The absence of pitch-based streaming

observed in experiment 1 suggests that impairment of stream segregation could be partially responsible for low speech perception performance in noise under CI simulation.

However, this result differs from those previously obtained in CI recipients using simple stimuli to test place-pitch or rate-pitch-based streaming (Hong and Turner, 2006; Chatterjee *et al.*, 2006). One explanation for this difference is that the complexity of spectral and temporal cues associated with even simple speech items such as vowels diminishes the availability of F_0 cues following CI processing. Because of the characteristics of speech, F_0 could not be reliably converted into place pitch or rate pitch, such as that associated with the previous experiments. Another possible reason is that the duration of the stimuli employed here was different from that used in previous studies. Hong and Turner (2006) and Chatterjee *et al.* (2006) used 60 and 50 ms stimuli, while the briefest vowels in the current study were 133 ms. A slower presentation rate can prevent obligatory streaming. However, to evaluate if obligatory streaming is involved in concurrent speech segregation in ecological situations, it is important to examine natural speech rates. Because streaming was not observed with speech rates matching those of natural speech in experiment 1 of the current study, it can be concluded that CI users may have difficulty taking advantage of streaming cues in cocktail party situations.

The current CI model employed numbers of channels similar to those found in modern CIs and an overall bandwidth somewhat smaller than that often employed clinically, but more appropriate for vowel reception. Because of this reduced overall bandwidth, and because the number of auditory channels available to a CI user can be somewhat lower than the number of physical electrodes (with the exception of successful "current steering" programming), the spectral detail provided by the current CI model may be even higher than that available in modern CIs. However, the comparison with CI user performance remains difficult because the noise-band vocoder simulation does not exactly mimic the perception of sounds by CI users. There is a long list of patient variables that are associated with CI users, but absent from consideration when using simulations. Also, CI recipients have generally experienced their CI for months prior to experimentation while NH participants in the current experiments were trained on noise-band vocoders for a few hours. Notably, the noise-band vocoder does not exactly mimic stimulation by the CI. A main difference is that the output of the vocoder is acoustic and then subjected to peripheral processing by the ear, while the CI involves direct electric stimulation.

Another effect that potentially reduces the strength of temporal pitch in the vocoder has been suggested by Hanna (1992). In the noise-band vocoder, the analysis filter bank and the resynthesis filter bank are typically the same. Thus, the noise carrier after modulation with the temporal envelope is then filtered again to suppress the sidebands, i.e., the modulation products that fall outside the band. Modulation depth is reduced by this resynthesis filtering for the narrowest bands, and temporal-pitch cues are then weakened. To fully preserve the voicing modulation, the bandwidth must be greater than twice the F_0. Using this metric, temporal-pitch cues were intact in the current vocoder channels beyond No. 6 in the Q_{12} condition and beyond No. 14 in the Q_{20} condition. In addition, as described by Hanna (1992), the peripheral filters of the normal ear play a role similar to that of the resynthesis filters of the vocoder, weakening again the modulation depth. Thus, although subjects were probably able to perceive temporal-pitch cues, their depth was reduced in the lower bands of the vocoder. In CI processors, neither resynthesis nor peripheral filtering occurs and temporal-pitch cues are not degraded in this way. Indeed, Laneau *et al.* (2006) observed that CI users had better F_0 discrimination abilities than NH listeners hearing noise-band vocoders.

Although it was found that temporal-pitch cues were not sufficient to produce obligatory streaming of vowel sequences, it cannot be entirely ruled out that pitch-based obligatory streaming can occur in CI users. In particular, it was found that an F_0-related spectral cue may have induced obligatory streaming. This spectral cue is potentially related to formant definition and was then not present in simpler stimulation. The current findings then suggest that these cues could potentially be available to CI listeners. Many attempts have been made in the past few years to provide a better encoding of pitch for CI recipients. Increasing the number of bands in the low frequencies better captures the first harmonic and can improve the perception of pitch (Carroll and Zeng, 2007). Unfortunately, with a fixed number of channels, increasing the number of bands in the low frequency region leads to a decrease in the number of bands in the higher regions, which appears to be detrimental for speech intelligibility. Hence, there seems to be a trade-off between pitch perception and speech intelligibility. Moreover, despite the increasing number of channels in CIs (up to 22), it seems that most CI recipients do not show a speech reception benefit from more than seven bands (Friesen *et al.*, 2001). These results suggest that pitch cues should be enhanced in existing bands to avoid the degradation of spectral cues required for speech perception. Instead of increasing spectral-pitch cues, Green *et al.* (2005) have enhanced temporal-pitch cues by adding to standard processing 100% AM at the F_0 frequency in all channels This manipulation improved the perception of prosody. However, again, it appeared that the modified processing had a detrimental effect on vowel recognition. These two strategies to enhance pitch perception in CI users do not account for the F_0-related spectral cues found in the current study. Further investigation is then required to evaluate the ability of CI users to take advantage of these cues for segregation of speech in ecological situations.

V. CONCLUSIONS

(1) Temporal-pitch cues available from noise-band vocoder simulations of a CI are not sufficient to induce obligatory streaming of speech materials at realistic speech rates.
(2) In contrast, the quantization of the spectrum in the vocoder enhances an F_0-related spectral cue that is capable of inducing streaming. This cue might be available to CI users.
(3) The use of temporal periodicity cues to induce obliga-

tory streaming in CI users is unlikely, unless it is assumed that these cues are stronger in actual CIs than in CI simulations.

ACKNOWLEDGMENTS

The authors wish to thank Andrew J. Oxenham and Christophe Micheyl for their helpful comments on this study. The authors would also like to thank two anonymous reviewers for helpful comments on a previous version of this manuscript. This work was supported in part by NIH Grant No. DC08594, by a doctoral grant from the Région Rhône-Alpes (France), and by Grant No. JC6007 from the Ministère de l'Enseignement Supérieur et de la Recherche (France). Manuscript preparation was supported in part by UK Medical Research Council Grant No. G9900369.

[1]Perceptual distance between vowels was calculated as the weighted Euclidian distance (in barks) in the F1–F'2 space, where F1 is the frequency of the first formant and F'2 is the frequency of the *effective second formant* defined as the weighted sum of the second to fourth formants (de Boer, 2000). Perceptual distance for a given sequence was then the sum of the perceptual distances between consecutive vowels.

American National Standards Institute (**1995**). *ANSI S3.7-R2003: Methods for Coupler Calibration of Earphones* (American National Standard Institute, New York).

American National Standards Institute (**2004**). *ANSI S3.21-2004: Methods for Manual Pure-Tone Threshold Audiometry* (American National Standard Institute, New York).

Baer, T., and Moore, B. C. J. (**1993**). "Effects of spectral smearing on the intelligibility of sentences in noise," J. Acoust. Soc. Am. **94**, 1229–1241.

Beauvois, M. W., and Meddis, R. (**1996**). "Computer simulation of auditory stream segregation in alternating-tone sequences," J. Acoust. Soc. Am. **99**, 2270–2280.

Berkes, P., and Zito, T. (**2007**). "Modular toolkit for data processing, version 2.1," http://mdp-toolkit.sourceforge.net

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge, MA).

Bregman, A. S. (**1978**). "Auditory streaming is cumulative," J. Exp. Psychol. Hum. Percept. Perform. **4**, 380–387.

Bregman, A. S., and Campbell, J. (**1971**). "Primary auditory stream segregation and perception of order in rapid sequences of tones," J. Exp. Psychol. **89**, 244–249.

de Boer, B. (**2000**). "Self-organization in vowel systems," J. Phonetics **28**, 441–465.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

Burns, E. M., and Viemeister, N. F. (**1981**). "Played-again SAM: Further observations on the pitch of amplitude-modulated noise," J. Acoust. Soc. Am. **70**, 1655–1660.

Burns, E. M., and Viemeister, N. F. (**1976**). "Nonspectral pitch," J. Acoust. Soc. Am. **60**, 863–869.

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.

Carroll, J., and Zeng, F. (**2007**). "Fundamental frequency discrimination and speech perception in noise in cochlear implant simulations," Hear. Res. **231**, 42–53.

Chatterjee, M., Sarampalis, A., and Oba, S. I. (**2006**). "Auditory stream segregation with cochlear implants: A preliminary report," Hear. Res. **222**, 100–107.

Cooper, H. R., and Roberts, B. (**2007**). "Auditory stream segregation of tone sequences in cochlear implant listeners," Hear. Res. **225**, 11–24.

Deeks, J. M., and Carlyon, R. P. (**2004**). "Simulations of cochlear implant hearing using filtered harmonic complexes: Implications for concurrent sound segregation," J. Acoust. Soc. Am. **115**, 1736–1746.

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 number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2410.

Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (**1998**). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels," J. Acoust. Soc. Am. **104**, 3583–3585.

Dudley, H. (**1939**). "The automatic synthesis of speech," Proc. Natl. Acad. Sci. U.S.A. **25**, 377–383.

Elhilali, M., and Shamma, S. (**2007**). in *Hearing: From Sensory Processing to Perception*, edited by B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey (Springer-Verlag, Berlin).

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

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.

Geurts, L., and Wouters, J. (**2001**). "Coding of the fundamental frequency in continuous interleaved sampling processors for cochlear implants," J. Acoust. Soc. Am. **109**, 713–726.

Green, T., Faulkner, A., Rosen, S., and Macherey, O. (**2005**). "Enhancement of temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification," J. Acoust. Soc. Am. **118**, 375–385.

Grimault, N., Bacon, S. P., and Micheyl, C. (**2002**). "Auditory stream segregation on the basis of amplitude modulation rate," J. Acoust. Soc. Am. **111**, 1340–1348.

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., and Collet, L. (**2001**). "Perceptual auditory stream segregation of sequences of complex sounds in subjects with normal and impaired hearing," Br. J. Audiol. **35**, 173–182.

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., and Collet, L. (**2000**). "Influence of peripheral resolvability on the perceptual segregation of harmonic tones differing in fundamental frequency," J. Acoust. Soc. Am. **108**, 263–271.

Grose, J. H., and Hall, J. W. (**1996**). "Perceptual organization of sequential stimuli in listeners with cochlear hearing loss," J. Speech Hear. Res. **39**, 1149–1158.

Hanna, T. E. (**1992**). "Discrimination and identification of modulation rate using a noise carrier," J. Acoust. Soc. Am. **91**, 2122–2128.

Hartmann, W. M., and Johnson, D. (**1991**). "Stream segregation and peripheral channeling," Music Percept. **9**, 115–184.

Healy, E. W., and Bacon, S. P. (**2002**). "Across-frequency comparison of temporal speech information by listeners with normal and impaired hearing," J. Speech Lang. Hear. Res. **45**, 1262–1275.

Healy, E. W., and Steinbach, H. M. (**2007**). "The effect of smoothing filter slope and spectral frequency on temporal speech information," J. Acoust. Soc. Am. **121**, 1177–1181.

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.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.

Laneau, J., Moonen, M., and Wouters, J. (**2006**). "Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants," J. Acoust. Soc. Am. **119**, 491–506.

Mackersie, C., Prida, T., 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.

McCabe, S. L., and Denham, M. J. (**1997**). "A model of auditory streaming," J. Acoust. Soc. Am. **101**, 1611.

Moore, B. C. J. (**1998**). *Cochlear Hearing Loss* (Whurr, London).

Moore, B. C. J., and Carlyon, R. P. (**2005**). in *Pitch: Neural Coding and Perception*, edited by C. J. Plack, A. J. Oxenham, R. R. Fay, and A. N. Popper (Springer, New York).

Moore, B.C.J., and Gockel, H. (**2002**). "Factors influencing sequential stream segregation," Acta. Acust. Acust. **88**, 320–332.

Moore, B. C. J., and Peters, R. W. (**1992**). "Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency

selectivity," J. Acoust. Soc. Am. **91**, 2881–2893.
Nooteboom, S. G., Brokx, J. P. L., and de Rooij, J. J. (**1978**). in *Studies in the Perception of Language*, edited by W. J. M. Level and G. B. F. d'Arcais (Wiley, New York), pp. 75–107.
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.
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.
Roberts, B., Glasberg, B. R., and Moore, B. C. J. (**2002**). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband," J. Acoust. Soc. Am. **112**, 2074–2085.
Rogers, C. F., Healy, E. W., and Montgomery, A. A. (**2006**). "Sensitivity to isolated and concurrent intensity and fundamental frequency increments by cochlear implant users under natural listening conditions," J. Acoust. Soc. Am. **119**, 2276–2287.
Rose, M. M., and Moore, B. C. J. (**2005**). "The relationship between stream segregation and frequency discrimination in normally hearing and hearing-impaired subjects," Hear. Res. **204**, 16–28.
Rose, M. M., and Moore, B. C. J. (**1997**). "Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **102**, 1768–1778.
Shannon, R. V. (**1983**). "Multichannel electrical stimulation of the auditory nerve in man. I: Basic psychophysics," Hear. Res. **11**, 157–189.
Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Stainsby, T. H., Moore, B. C. J., and Glasberg, B. R. (**2004**). "Auditory streaming based on temporal structure in hearing-impaired listeners," Hear. Res. **192**, 119–130.
Stickney, G. S., Assmann, P. F., Chang, J., and Zeng, F. (**2007**). "Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences," J. Acoust. Soc. Am. **122**, 1069–1078.
Stickney, G. S., Zeng, F., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.
Tong, Y. C., and Clark, G. M. (**1985**). "Absolute identification of electric pulse rates and electrode positions by cochlear implant patients," J. Acoust. Soc. Am. **77**, 1881–1888.
Townshend, B., Cotter, N., Compernolle, D. V., and White, R. L. (**1987**). "Pitch perception by cochlear implant subjects," J. Acoust. Soc. Am. **82**, 106–115.
Tsuzaki, M., Takeshima, C., Irino, T., and Patterson, R. D. (**2007**). in *Hearing: From Sensory Processing to Perception*, edited by B. Kollmeier, G. Klump, V.Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey (Springer-Verlag, Berlin).
van Noorden, L. P. A. S. (**1975**). "Temporal coherence in the perception of tones sequences," Ph.D. thesis, Eindhoven University of Technology, The Netherlands.
Vliegen, J., Moore, B. C. J., and Oxenham, A. J. (**1999**). "The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task," J. Acoust. Soc. Am. **106**, 938–945.
Vliegen, J., and Oxenham, A. J. (**1999**). "Sequential stream segregation in the absence of spectral cues," J. Acoust. Soc. Am. **105**, 339–346.
Zeng, F. G. (**2002**). "Temporal pitch in electric hearing," Hear. Res. **174**, 101–106.