Sequential stream segregation of voiced and unvoiced speech sounds based on fundamental frequency

Marion A. David, Mathieu Lavandier, Nicolas Grimault, Andrew J. Oxenham

Running title: Segregation of speech sounds based on F0 differences

Sequential stream segregation of voiced and unvoiced speech sounds based on fundamental frequency

Marion David [a], Mathieu Lavandier [b], Nicolas Grimault [c] and Andrew J. Oxenham [a]

[a] *Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455, USA*

[b] *Univ. Lyon, ENTPE, Laboratoire Génie Civil et Bâtiment, Rue M. Audin, F-69518 Vaulx-en-Velin Cedex, FRANCE*

[c] *Cognition Auditive et Psychoacoustique, Centre de Recherche en Neurosciences de Lyon, Université Lyon 1, UMR CRNS 5292, Avenue Tony Garnier, 69366 Lyon Cedex 07, FRANCE*

13 **Abstract**

14 Differences in fundamental frequency (F0) between voiced sounds are known to be a strong cue
15 for stream segregation. However, speech consists of both voiced and unvoiced sounds, and less
16 is known about whether and how the unvoiced portions are segregated. This study measured
17 listeners' ability to integrate or segregate sequences of consonant-vowel tokens, comprising a
18 voiceless fricative and a vowel, as a function of the F0 difference between interleaved sequences
19 of tokens. A performance-based measure was used, in which listeners detected the presence of a
20 repeated token either within one sequence or between the two sequences (measures of voluntary
21 and obligatory streaming, respectively). The results showed a systematic increase of voluntary
22 stream segregation as the F0 difference between the two interleaved sequences increased from 0
23 to 13 semitones, suggesting that F0 differences allowed listeners to segregate speech sounds,
24 including the unvoiced portions. In contrast to the consistent effects of voluntary streaming, the
25 trend towards obligatory stream segregation at large F0 differences failed to reach significance.
26 Listeners were no longer able to perform the voluntary-streaming task reliably when the
27 unvoiced portions were removed from the stimuli, suggesting that the unvoiced portions were
28 used and correctly segregated in the original task. The results demonstrate that streaming based
29 on F0 differences occurs for natural speech sounds, and that unvoiced portions are correctly
30 assigned to corresponding voiced portions of the speech sounds.

31 *Keywords*: Stream segregation, Fundamental frequency, Speech sounds

1. Introduction

Speech intelligibility in complex auditory environments, such as a cocktail party (Cherry, 1953), relies on our natural ability to perceptually segregate competing voices. To be intelligible, the sequence of sounds spoken by each person must be integrated into a single perceptual stream, and must be segregated from the speech sounds produced by other people. Auditory stream segregation and integration have been studied using both speech and non-speech sounds.

A large body of literature has documented the cues by which simple (non-speech) sounds are perceptually integrated and segregated (e.g., Bregman, 1990; Moore and Gockel, 2002, 2012). One important segregation cue involves differences in frequency or fundamental frequency (F0) between pure tones (Miller 1957; van Noorden 1975) and complex tones (Vliegen and Oxenham, 1999), respectively. One difficulty with generalizing the results from studies of streaming to real-world listening is that streaming studies often use sequences of sounds that are exact repetitions of each other, without the variations that are common in everyday situations. Some exceptions include studies of melody discrimination (e.g., Hartmann and Johnson, 1991), and a study involving two interleaved sequences of vowels that differed in F0 (Gaudrain et al. 2007). Listeners in that study were asked to report the order of presentation of the vowels either between or within the two interleaved sequences. Performance in the between-sequence task decreased significantly, while performance in the within-sequence task improved significantly, as the difference in F0 (ΔF0) between the two streams increased. Although this result shows that sequential voiced speech sounds can be segregated based on F0 differences, real speech also includes many unvoiced sounds, such as fricatives, which must be assigned to the correct speaker and segregated from other competing sounds.

Numerous studies of speech perception in the presence of competing speech have shown that F0 and intonation differences between a target and an interfering speaker can indeed improve the intelligibility of a target (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; Bird and Darwin, 1998; Darwin et al., 2003), along with other cues, such as differences in vocal tract length (Darwin and Hukin, 2000; Darwin et al., 2003; Gaudrain and Başkent, 2015) or intensity differences (Brungart, 2001). However, these measures were based on sentence intelligibility. Because of the numerous linguistic and other context effects present in speech, such stimuli do not provide a strong test of whether all voiced and unvoiced segments are correctly assigned to the correct speaker, as some degree of reconstruction could occur based on linguistic or lexical context and constraints.

A stronger test of the binding between consonants and vowels was provided by Cole and Cole and Scott (1973), who studied the perceptual organization of repeating syllables consisting of an unvoiced fricative consonant and a voiced vowel (CV), all with the same vowel (/a/) but with different consonants. They found that listeners' ability to judge the order of the sounds was best when the natural sounds were presented, and worsened if the formant transitions between the consonant and its vowel were removed from the vowels. They argued that these vowel transitions play an important role in binding adjacent segments of speech. A more recent study (Stachurski et al., 2015) used the verbal transformation effect (Warren, 1961) to determine the extent to which formant transitions bind vowels to their preceding consonant. Stachurski et al. (2015) found that the number of verbal transformations reported decreased when the formant transitions were left intact, suggesting that the transitions provided additional binding between the consonant and its following vowel, particularly when the formant transition itself was more pronounced.

77 Although these studies suggest that formant transitions assist in binding successive consonant
78 and vowel pairs, none of them has studied the extent to which this binding is maintained in the
79 presence of competing streams, as would be encountered in a multi-talker environment. The
80 purpose of the present study was to test whether successful streaming of interleaved sequences of
81 speech sounds can be achieved based solely on differences in F0 between the voiced portions of
82 speech, and thus whether the unvoiced segments can be segregated into the correct streams by
83 virtue of their companion voiced segments. On the one hand, the temporal proximity of the
84 unvoiced and voiced portions of a CV pair, along with the formant transitions, might assist in the
85 perceptual fusion of the unvoiced and voiced portions (Cole and Scott, 1973; Stachurski et al.,
86 2015). On the other hand, repeating sequences of spectrally dissimilar sounds (such as the
87 fricative consonant and vowel) can lead to perceptual segregation and, in some cases, spurious
88 perceptual organization (Harris, 1958), even when formant transitions are maintained (Stachurski
89 et al., 2015). Here, naturally spoken CV pairs were generated to produce speech sounds that
90 contained both unvoiced and voiced segments. Sequences of speech sounds were then generated
91 by concatenating the speech sounds in random order into sequences. Two such sequences were
92 temporally interleaved, and a difference in F0 was introduced between the interleaved sequences
93 to produce a pattern of speech tokens with alternating F0, and thus induce stream segregation.
94 Performance was measured in tasks that either favored perceptual integration of all the sounds
95 into a single stream or favored perceptual segregation of the alternating sounds into two separate
96 streams.

2. Experiment 1: Within- and across-sequence repetition detection with consonant-vowel pairs

2.1 Rationale

The aim of this experiment was to test whether sequential stream segregation of CV tokens can be elicited by differences in F0 between the voiced portions of the tokens. Voiceless fricatives were used as consonants to provide noise-like aperiodic stimuli that did not carry F0 information. Therefore, successful streaming based solely on F0 differences would require additional binding of the voiced and voiceless segments of each CV token. Such binding can occur in naturally uttered speech signals due to spectral transitions between the consonant and vowel (Cole and Scott, 1973; Stachurski et al., 2015). The present experiment tests whether such binding is sufficient to allow segregation of competing streams.

2.2 Methods

2.2.1. Stimuli

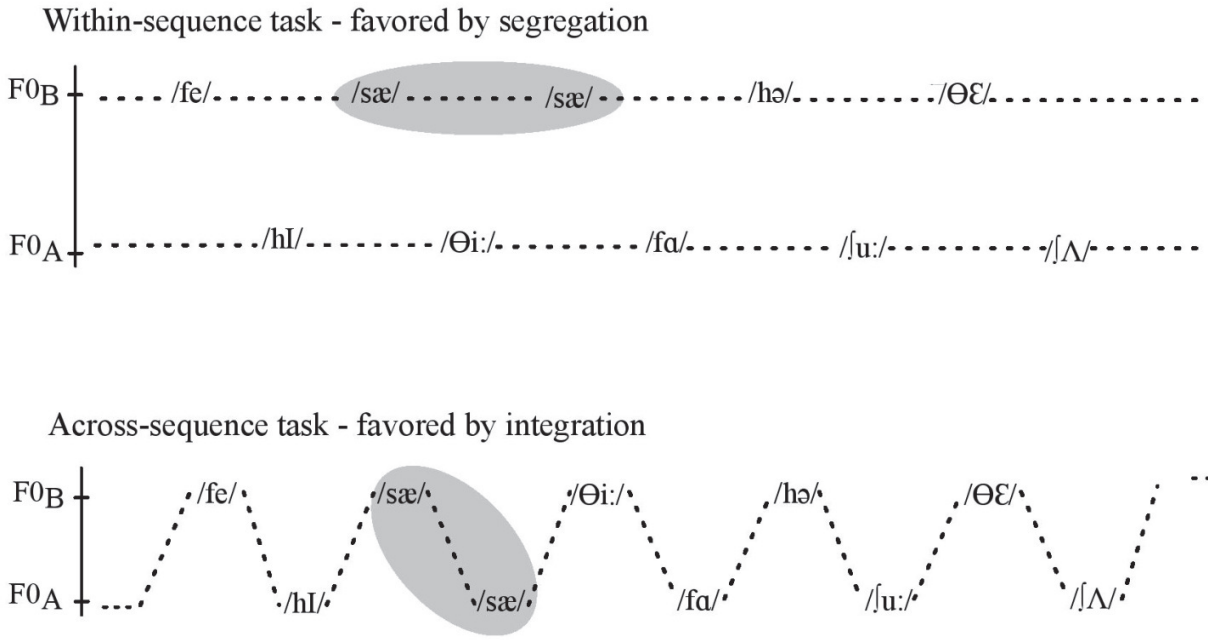
The speech sounds were naturally uttered pairs of voiceless fricative consonants and voiced vowels. Because the consonant-vowel stimuli were recorded as a whole, they included a fricative part (the consonant), a transition part (the vocalic part still containing some consonant information) and a voiced part (the vowel). A set of 45 such sounds were recorded by two speakers, one male and one female, both of whom were native speakers of American English. The recordings were made with a microphone (Sennheiser E914) and portable digital recorder (Marantz PMD670) in a sound attenuating booth. The stimulus set was composed of five voiceless fricative consonants ([f], [s], [□], [ʃ] and [h]) combined with nine vowels ([æ], [e], [iː], [I], [ɘ], [□], [□], [ɑ] and [uː]). The [h] is not often considered in studies investigating fricative

consonants (Jongman et al., 2000); however, [h] is defined as a glottal fricative consonant in the International Phonetic Alphabet (IPA), and so was included here.

The stimuli had to be short enough to produce automatic or obligatory stream segregation (van Noorden, 1975), but long enough to contain information from both the consonant and vowel. The duration of each token was therefore limited to 160 ms, with 40 ms inter-token intervals, leading to an onset-to-onset time of 200 ms which is close to the upper limit for observing obligatory stream segregation (van Noorden, 1975; Micheyl and Oxenham, 2010a; David et al, 2015) . The beginning and end of the recorded speech sounds were truncated and gated on and off with 10-ms raised-cosine ramps. The truncation points were chosen manually to ensure that the consonant and vowel parts of the stimulus had approximately the same length. The spectral shapes of the different vowels were, of course, different, but the spectral shape of the steady-state portion of each vowel did not differ much in the context of different consonants, as expected. The pitch contours of the tokens were flattened using Praat software (Boersma and Weenink, 2001).The stimuli were then resynthesized using a pitch synchronous overlap-add technique (PSOLA), widely used for F0 manipulations of speech sounds, which has minimal effect on the spectral shape of the CV tokens, including the vocalic portions.

Listeners were presented with interleaved sequences in an ABAB… format, with the A and B sequences presented at different F0s. There were 14 speech tokens in each of the A and B sequences, for a total of 28 speech tokens in each presentation, with the speech tokens selected randomly (without replacement) from the total set of 45 tokens for each presentation. The F0 of the A tokens was constant at 110 Hz and 220 Hz for the male and female voice, respectively, while the F0 of the B tokens was set to be ΔF0 semitones above the F0 of A (0, 1, 3, 5, 7 and 9 semitones, i.e., approximately 110, 116, 131, 147, 165, 185 Hz, and 220, 233, 262, 294, 330, 370

142 Hz for the male and female voice, respectively). In half the presentations, selected at random, a
143 consecutive repetition of a CV token was introduced. Depending on the condition (within- or
144 across-sequence task), the repetition occurred in one of the sequence (as two consecutive As) or
145 across the sequences (as a consecutive A and B), as shown in Fig. 1. In the within-sequence task,
146 the listeners were asked to attend to the voice with the lower pitch (i.e., the A sequence). No
147 repetitions were introduced in the higher-F0 sequence (B). In the across-sequence task, listeners
148 were instructed to attend to the entire interleaved ABAB… sequence. The repetition was
149 introduced at a random position sometime after the 12th token, in order to allow time for the
150 build-up of segregation (Anstis and Saida, 1985; Haywood and Roberts, 2010). Performance was
151 predicted to be best in the within-sequence task when listeners were able to segregate the
152 interleaved sequence into two streams and so to hear out a repetition within one stream without
153 interference from the other stream, and to be best in the across-sequence task when listeners were
154 able to integrate the sequence into one single stream, and so detect a repetition of a CV that
155 occurred across the two sequences. Listeners are typically able to judge accurately the relative
156 timing of consecutive tokens only when they fall within a single stream (Roberts et al., 2002;
157 Micheyl and Oxenham, 2010).
158



159

160 **Fig. 1:** Structure of the interleaved sequences in the within-sequence (top panel) and across-
161 sequence (bottom panel) tasks. The syllables within a shaded region correspond to a repeated
162 token. In half the presentations, the interleaved sequences consisted of only different stimuli (not
163 shown) and in the other half, a repeat was introduced. In the within-sequence task, performance
164 should improve when the sequences are heard as two separate streams, whereas in the across-
165 sequence task, performance should improve when the interleaved sequences are heard as a single
166 stream.

167 [FIG. 1 ABOUT HERE]

168 *2.2.2. Procedure*

169 In both tasks, listeners had to indicate whether or not the interleaved sequence contained a
170 repeat. Feedback was provided after each response. Listeners' sensitivity to the repetition (*d'*)
171 was estimated by taking the inverse cumulative normal distribution function (z-transform) of the
172 hit rate (H, i.e., proportion of repeats correctly detected) and subtracting from that the same
173 transformation of the false alarm rate (FA, i.e., proportion of repeats reported in trials with no

174 repeats), with a correction for 100% or 0% H or FA rates by using 1-1/(2*N*) and 1/(2*N*),
175 respectively, where *N* is the total number of trials (Macmillan and Creelman, 2004).

176 The experiment involved two sessions, with each session devoted to one of the two tasks.
177 Half the listeners started with the across-sequence task and the other half started with the within-
178 sequence task. In each session, the listeners completed thirteen runs per talker (i.e., 26 runs in
179 total). For each run, 2 repetitions of the 12 conditions (6 values of ΔF0, each with repeat and no-
180 repeat conditions) were completed, resulting in a total of 624 sequences tested for each task.
181 Both sessions took place in a sound-attenuating booth. Stimulus presentation and response
182 collection were controlled using the AFC software package (Ewert, 2013) under MATLAB
183 (Mathworks, Natick, MA). The stimuli were presented diotically at 65 dB SPL via HD 650
184 headphones (Sennheiser, Wedemark, Germany).

185

186 *2.2.3 Listeners*

187 Sixteen listeners were recruited for this experiment. All of them were native speakers of
188 American English. They all had normal hearing (i.e., pure tone threshold of less than 20 dB HL
189 at octave frequencies between 200 and 8000 Hz), and were paid an hourly wage for their
190 participation. In addition to screening for normal hearing, a selection criterion was used to ensure
191 that each listener was able to perform the task. Each subject's performance in each of the 24
192 conditions (two talkers, two tasks, and six values of ΔF0) was calculated in terms of d' and a
193 one-sample two-tailed t-test was performed to determine whether the average performance of
194 each subject was significantly different from chance ($d' = 0$). All the listeners whose overall
195 performance, pooled across all conditions, was significantly different from chance were included

196 in the analyses. One of the listeners did not perform above chance using this test, and so their
197 data were excluded from further consideration. The remaining 15 listeners were aged between 18
198 and 24 years (seven females, eight males, average age = 19.5 years, standard deviation, SD = 1.6
199 years).

200

201 *2.3. Results*

202 The *d′* scores in each of the two tasks were subjected to a mixed-model analysis of variance
203 (ANOVA), with the order of the tasks (within- or across-sequence condition first) as a between-
204 subjects factor, and speaker gender (male/female) and ΔF0 (1-9 semitones) as within-subjects
205 factors. Neither the main effects of speaker gender and task order nor their interactions were
206 significant ($p > 0.2$ in all cases). For this reason, the results shown in Fig. 2 are averaged across
207 participants, speaker gender, and task order. The left panel shows the results for the within-
208 sequence task and the right panel shows the results for the across-sequence task. The main effect
209 of ΔF0 was significant in the within-sequence task [$F(1,14) = 24.4$, $p < 0.001$], with a significant
210 linear trend ($p = 0.003$), reflecting a systematic increase in performance with increasing ΔF0, as
211 would be expected if an increase in the F0 separation led to improved segregation between the
212 two interleaved sequences, making it easier for subjects to attend selectively to one sequence (the
213 one with the lower pitch) to detect the repetition. In the across-sequence condition (right panel),
214 the main effect of ΔF0 was not significant [$F(1,14) = 1.88$. $p = 0.183$]. It appears, therefore, that
215 introducing an F0 difference of up to 9 semitones between the two interleaved sequences did not
216 result in obligatory streaming, or in the inability to detect patterns that occurred between the two
217 sequences.

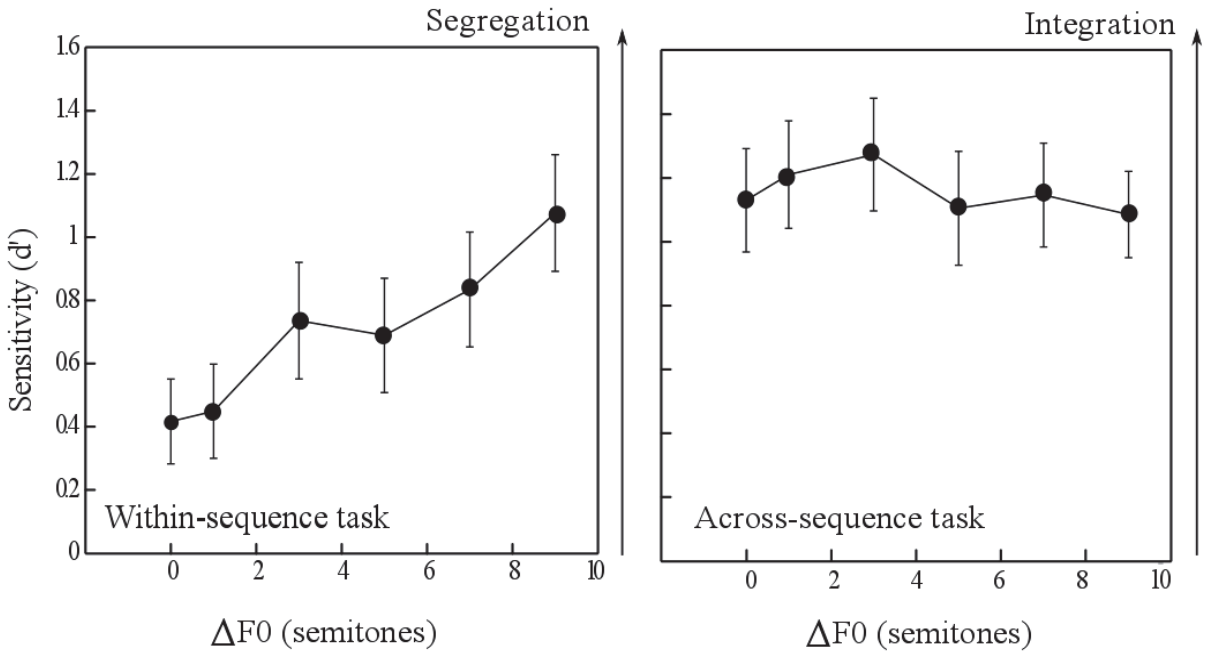


Fig. 2: Mean performance across fifteen listeners for the within- (left panel) and the across-(right panel) sequence tasks in Experiment 1. In the within-sequence task, high *d′* values at large ΔF0 values indicates a greater tendency to segregate the sequences into two streams; in the across-sequence task, the generally high *d′* values indicate an ability to integrate the interleaved sequence into a single stream, despite the F0 difference between the two sequences. The error bars represent ±1 standard error of the mean.

[FIG 2 ABOUT HERE]

2.4. Discussion

Listeners were able to make use of a difference in F0 between the two sequences of speech sounds in order to detect a repeated speech token within one of the sequences. This result is in agreement with previous studies, which found that stream segregation can be elicited by a difference in F0 when listeners attempt to segregate sounds (Darwin et al., 2003; Gaudrain et al.

2007). This improvement occurred despite the fact that the speech sounds contained voiceless as well as voiced elements, meaning that the F0 cues were only salient for a portion of the speech sounds. One interpretation of this outcome is that listeners were able to perceptually fuse the voiceless and voiced parts of each speech sound even without the F0 cue in the consonant part. Another possibility, however, is that listeners attended only to the voiced part of the speech sounds and responded based only on those parts.

In the case where listeners had to detect a repetition across sequences, there was little evidence for a worsening in performance with increasing F0 difference, as would have been expected based on streaming considerations. Again, multiple explanations are possible. First, a shallower slope than for the within- sequence task is expected, based on the fact that listeners were attempting to segregate in the within-sequence task, and to integrate in the across-sequence task. Indeed, given the definitions proposed by van Noorden (1975), the thresholds of obligatory and voluntary stream segregation correspond to the temporal coherence and fission boundaries (TCB and FB), respectively. Since the TCB requires larger stimulus dissimilarity for streaming to occur compared to FB, obligatory stream segregation was expected to be less affected by a difference in F0 than voluntary stream segregation. Second, the relatively long onset-to-onset time of 200 ms provides only a weak impetus for obligatory stream segregation (van Noorden, 1975). Third, broadband sounds that overlap in spectrum do not always produce an obligatory streaming effect. For instance, Vliegen et al. (1999) found that larger differences in F0 than were tested here were necessary to induce obligatory segregation of sequences of complex tones with overlapping harmonic spectra. Fourth, it is possible that listeners were simply detecting a repeat in the voiceless portions of the speech sounds. In this case, introducing an F0 difference would not necessarily worsen performance in the across-stream task, as the voiceless portions may not

256 have been segregated. Experiment 2 attempts to distinguish between these alternative
257 explanations.

258

259 ## 3. Experiment 2: Separate contributions of vowels and consonants to repetition
260 detection

261 ### *3.1. Rationale*

262 Experiment 1 showed that F0 differences seemed to allow listeners to segregate sequences of
263 speech sounds that contained both voiced and unvoiced information. However, the repetition of
264 one token could have been detected by either the repetition of just the vowel or just the
265 consonant. To test whether listeners were indeed streaming both the vowels and consonants, this
266 experiment ensured that all the non-target trials, which did not contain a repeated CV, instead
267 contained a repetition of either the consonant or the vowel. In this way, good performance would
268 only be possible if the listener was able to perceive the repetition of both the consonant and the
269 vowel. In addition, a larger maximum F0 separation was achieved without resorting to an
270 unnatural combination of F0 and vocal tract length, by increasing the F0 of the higher stream and
271 decreasing the F0 of the lower stream, so that neither stream was more than six semitones away
272 from its original F0.

273

3.2. Method

3.2.1. Stimuli

The stimulus tokens used in this experiment were the same as those in Experiment 1. However, because the speaker's gender was found to have no effect, only the male voice was used here. To encourage attention to the entire interleaved sequence on each trial, the length of each sequence was randomized to be between 16 and 28 tokens long. The repeat (if present) was always presented in the penultimate pair of tokens.

To allow us to test a wider range of ΔF0 values, the F0s of the A and B tokens were varied, with the F0 of the A tokens decreasing and the F0 of the B tokens increasing. The values of ΔF0 tested were 0 semitones ($F0_A$ = 110 Hz, $F0_B$ = 110 Hz), 3 semitones (104 and 123 Hz), 5 semitones (98 and 131 Hz), 7 semitones (92 and 139 Hz), 9 semitones (87 and 147 Hz) and 13 semitones (78 and 165 Hz).

3.2.2. Procedure

To investigate whether the listeners' responses were based more on the vowels or the consonants, 50% of the presentations, selected at random, included a consecutive repetition of a full token (consonant and vowel, referred to as a "full repeat"), 25% of the presentations included a repetition of only the consonant, and 25% included a repetition of only the vowel (these last two cases being referred to as a "half repeat"). The hit rate (H) corresponded to the proportion of full repeats that were detected; the false alarm (FA) rate corresponded to the proportion of trials in which a repeat was reported when in fact only a half-repeat was presented. Because of the experiment's design, it was possible to calculate separately the FA for the

296 consonant-only and vowel-only repeats. As in Experiment 1, listeners were instructed to attend
297 to the low-pitch sequence (A sequence) in the within- sequence task, and no repeat was
298 introduced in the high-pitch sequence (B sequence).

299 The within- and across-sequence tasks were completed in a single two-hour session. Half the
300 listeners started with the across-sequence task and the other half started with the within-sequence
301 task. In each session, the listeners completed fifteen runs per task, for a total of 30 runs. For each
302 run, 24 conditions (6 values of ΔF0, each with 2 full repeats, 1 vowel-only repeat, and 1
303 consonant-only repeat) were presented, resulting in a total of 720 sequences tested. The
304 experimental setup was the same as for Experiment 1.

305

306 *3.2.3. Listeners*

307 The same selection criteria were used for listeners as in Experiment 1. Twenty-six out of
308 twenty-eight listeners tested, aged from 18 to 33 years (twelve females, fourteen males, average
309 age = 22.5 years, SD = 4.2 years), met the criterion of performing the task above chance on
310 average. One listener had already participated in Experiment 1. All the listeners were native
311 speakers of American English and were paid for their participation.

312

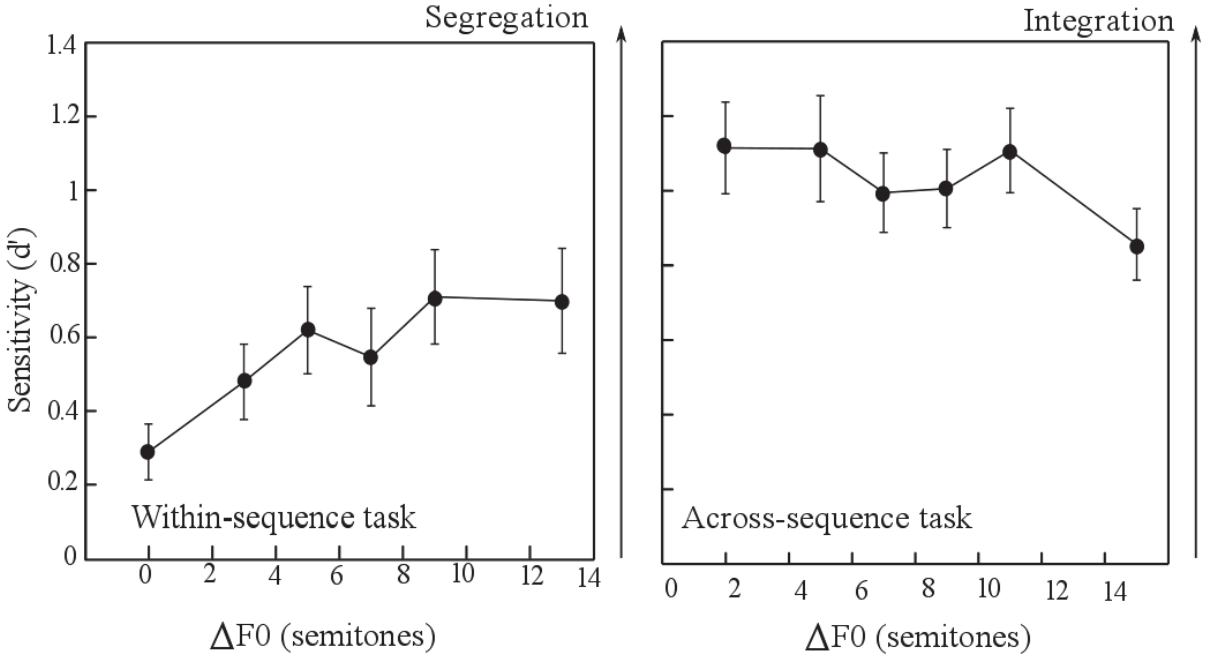
313 *3.3. Results*

314 The d' scores in each of the two tasks were subjected to a mixed-model ANOVA, with the
315 order of the tasks (within- or across-sequence condition first) as a between-subjects factor, and
316 ΔF0 (0-13 semitones) as a within-subjects factor. For both tasks, the effect of the order of the

317 tasks was not significant [$F(1,9) = 0.512$, $p = 0.482$ and $F(1,9) = 0.373$, $p = 0.548$, for the within
318 and across-sequence tasks, respectively]. Thus the mean data, averaged across listeners and task
319 orders, are shown in Fig. 3. The left and right panels correspond to the within- and across-
320 sequence tasks, respectively. As in Experiment 1, the main effect of ΔF0 was significant for the
321 within-sequence task [$F(1,25) = 12.57$, $p = 0.002$], with a significant linear trend ($p = 0.018$),
322 reflecting the improvement in performance with increasing ΔF0 (Fig. 3, left panel). Also in line
323 with Experiment 1, the effect of ΔF0 failed to reach significance for the across-stream task
324 [$F(1,25) = 3.31$, $p = 0.081$], although a trend was apparent for decreasing performance at the
325 very largest value of ΔF0.

326

327



328

329 **Fig. 3:** Mean performance across the twenty-six listeners who could perform the task in
330 terms of *d'* scores for the within- (left panel) and the across- (right panel) sequence tasks in
331 Experiment 2. In the within- sequence task, high *d'* values indicate a greater tendency to
332 segregate the sequences apart; in the across- sequence task, high *d'* values indicate a greater
333 tendency to integrate the interleaved sequence into one single stream. The error bars correspond
334 to ±1 standard error of the mean.

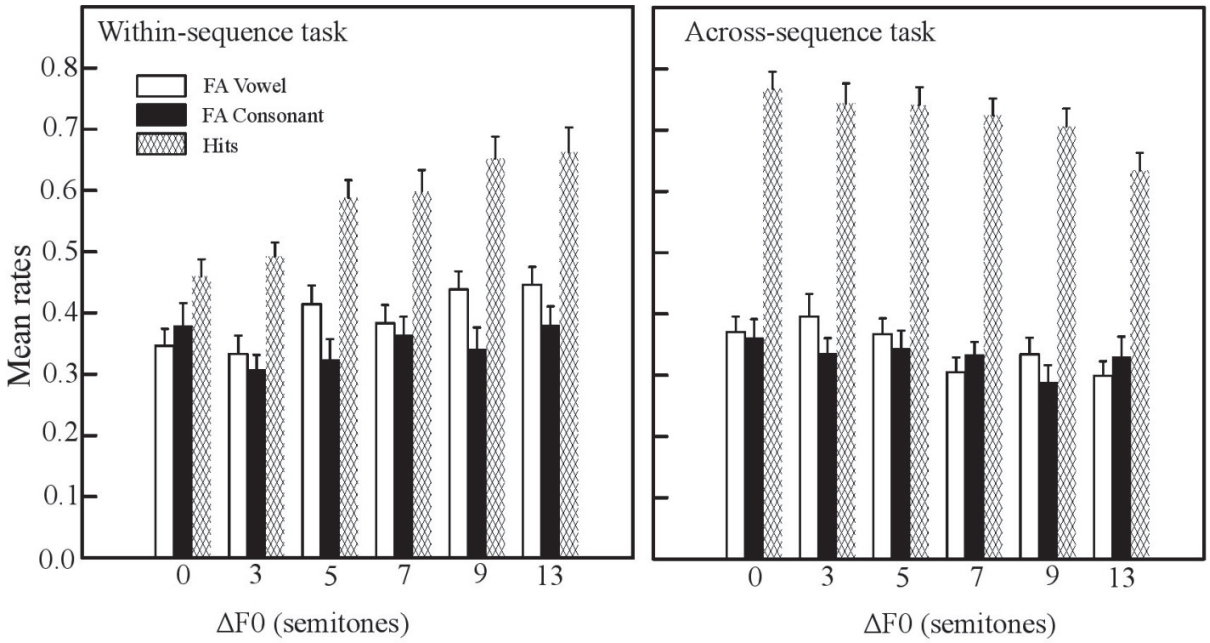
335 [FIG 3 ABOUT HERE]

336 In this experiment, the no-repeat trials included a repetition of either the consonant or the
337 vowel, but not both. To determine whether performance relied more on one speech segment than
338 the other, an analysis of the FA rates was carried out. The FA rates in response to a vowel-only
339 or a consonant-only repeat are shown in Fig. 4, along with the H rates. It can be seen that the FA
340 rates for the vowel-only and consonant-only repeat trials were quite similar. This outcome
341 suggests that performance was based not on just the vowels or just the consonants, but instead

342 that listeners were integrating information from the entire CV to perform the task. Nevertheless,
343 the FA rates associated with the vowels were slightly but consistently higher than the FA rates
344 associated with the consonants in the within- sequence task, in line with expectations given the
345 more salient information for streaming and identification present in the vowels.

346 A mixed-model ANOVA was performed on the FA rates for both tasks separately, again with
347 the order of the tasks (within- or across-sequence condition first) as a between-subjects factor,
348 and FA type (vowel or consonant) and ΔF0 (0-13 semitones) as a within-subjects factors. The
349 effect of the order of the task was not significant in the within- and across-sequence tasks, nor
350 the effect of the FA type. The effect of ΔF0 was significant [$F(1,9) = 6.83$, $p = 0.028$] in the
351 within-sequence task but not in the across-sequence task. None of the interactions were
352 significant in either task.

353



354

355 **Fig. 4:** Mean false- and H rates for the within- and across-sequence tasks (left and right panels,
356 respectively). The error bars correspond to ±1 standard error of the mean.

357 [FIG 4 ABOUT HERE]

358 *3.4. Discussion*

359 The listeners were able to detect a repetition introduced either across or within the sequences.
360 Segregation became significantly easier as ΔF0 increased. Although there was a trend for
361 integration to become more difficult with increasing ΔF0, it failed to reach significance.

362 The main purpose of this experiment was to test whether listeners were using the full CV,
363 rather than just the vowel or just the consonant, to perform the task. The fact that listeners were
364 able to perform the task at a similar level of performance as found in Experiment 1, despite the
365 fact that each trial had a repeat of either the vowel or the consonant, suggests that listeners could
366 indeed perceive and segregate the entire CV. The generally similar FA rates for both the vowel-

only and consonant-only trials suggest that both influence performance, although there was a tendency for the vowels to produce higher FA rates.

There remains another potential explanation for the outcomes of this experiment, which would not necessarily require the streaming of the unvoiced portions of the speech sounds: it may be that there is sufficient information regarding the identity of the consonant embedded in the voiced transition between the consonant and the vowel, due to effects of coarticulation (Harris, 1958; Repp, 1981; Wagner et al. , 2006). In other words, listeners may have relied solely on the voiced portions of the speech to segregate the sounds, but were able to derive the identity of the consonant from the initial portion of the vowel. This possibility was tested in Experiment 3.

4. Experiment 3: Testing for the presence of consonant information in the vowel

4.1. Rationale

The aim of Experiment 3 was to test the hypothesis that listeners were using the voiced portion of the CV to extract the identity of the consonant. If this were the case, then no conclusions can be drawn regarding the streaming of the unvoiced portions of the speech sounds. Whalen (1984) showed that a mismatched transition between the consonant and the vowel increased the reaction time for the identification of CV syllables without influencing the response accuracy. This result shows the importance of the fricative content in the transition part (i.e., the vocalic formant transition) on the identification in CV stimuli. It has also been shown that matched transitions are needed for non-sibilant fricative consonants (in the present case [f], [h] and [☒]) to ensure their correct identification (Harris, 1958), even if there is some variability

among listeners (Repp, 1981). The vocalic formant transition has been shown to have an influence on the perception and the identification of the unvoiced portion of a CV token (Wagner et al., 2006). To explore this possibility, this experiment tested listeners' ability to perform the task used in Experiment 2, but with the stimuli truncated to contain only the voiced portion of each CV pair. If listeners were able to still perform the task with the truncated stimuli, then it would suggest that segregation can be based solely on the voiced portions of the speech. On the other hand, if listeners are not able to perform the task with the truncated stimuli, that would suggest that listeners require the unvoiced portions to perform the task, and that these unvoiced portions are successfully segregated even without any F0 information.

4.2 Method

4.2.1 Stimuli

The harmonics-to-noise ratio (HNR) was evaluated for each stimulus used in Experiments 2. The HNR, which was initially used to define the degree of hoarseness (Yumoto, 1982), enables the evaluation of the relative weight of the noise and the harmonic content (in the present case the fricative consonant and vowel, respectively). The HNR was calculated over time-frame steps of 2 ms. This analysis of HNR over time revealed an inflection corresponding to the transition part between the consonant and vowel (see Fig. 5). The midpoint of the inflection was taken as the reference where the energy of the fricative consonant was roughly equivalent to the energy of the voiced vowel. Only the 80 ms of vowel following the midpoint (included the vocalic transition) was preserved and windowed with 5-ms raised-cosine onset and offset ramps. In this way, the stimuli contained both the vocalic formant transition and the vocalic part of the initial token, but not the unvoiced part of the fricative. The offset-to-onset time was increased from 40 ms to 120 ms, so that the onset-to-onset time remained at 200 ms.

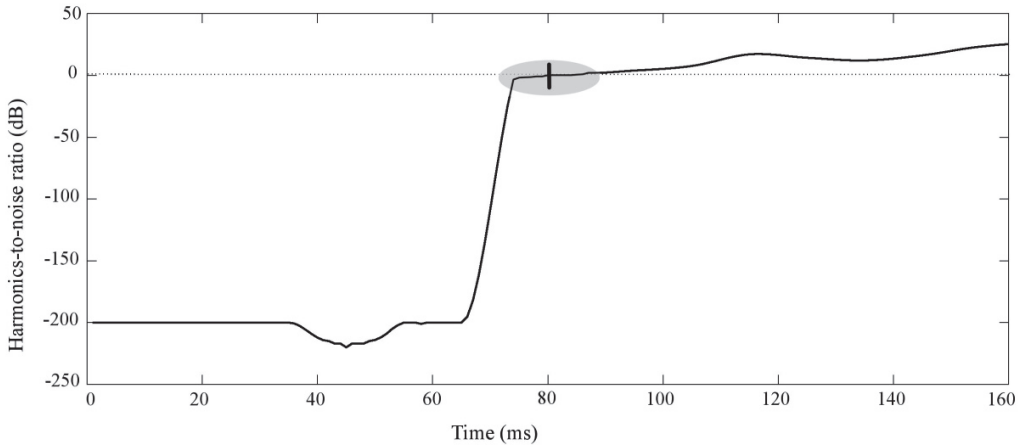


Fig. 5: An example of harmonics-to-noise ratio (HNR) as a function of time. When the HNR is less than 0 dB, the noise part (fricative part in our case) is dominant, and when the HNR is greater than 0 dB, the harmonic part (vowel part in our case) is dominant. The inflection, representing the transition, is displayed by the grey zone and the midpoint is indicated by the vertical bar.

[FIG 5 ABOUT HERE]

4.2.2. Procedure

The two tasks were the same as in the first two experiments (see Fig. 1) and the conditions were similar to those in Experiment 2. Half of the presentations presented a "full repeat" (both consonant and vowel repeated) and half of the presentations presented a "half repeat" (either consonant or vowel repeated). The same F0 differences of 0, 3, 5, 7, 9 and 13 semitones were tested.

As in Experiment 2, the two tasks were completed in a single two-hour session. Half of the participants started with the across-sequence task and the other half started with the within-sequence task. In each session, the listeners completed fifteen runs per task, for a total of 30 runs. For each run, 24 conditions (6 values of ΔF0 with 4 repeat types: 2 full repeats, 1 vowel-only

repeat and 1 consonant-only repeat) were presented, resulting in a total of 720 sequences tested. The experiment setup remained the same as in Experiments 1 and 2.

4.2.3. Listeners

Sixteen listeners took part in the experiment, aged from 18 to 58 years (eight females and eight males, average age = 24.4 years, SD = 9.9 years). All of them were native speakers of American English and had normal or near-normal hearing (one subject had a slight bilateral hearing loss at 8 kHz, with 35 and 20 dB HL in the right and left ear, respectively). They were paid an hourly wage for their participation. In the previous two experiments, listeners were required to perform above chance overall in order to be included in the analysis. However, in this experiment, only eight of the sixteen listeners would have achieved criterion performance. For this reason the results from all the listeners are shown below.

4.3. Results and Discussion

The question asked by this experiment was whether listeners could perform the task based only on the voiced segments of the speech stimuli. The fact that only eight of sixteen subjects passed the selection criterion (even without any correction for multiple statistical tests) suggests that listeners were generally *not* able to reliably perform the task. Confirming this expectation, Fig. 6 shows that overall performance, averaged across subjects, was also poor, with d' values not exceeding 0.3 in the within-sequence task and not exceeding 1 in the across-sequence task. In the within-sequence task (left panel), a repeated-measures ANOVA revealed no main effect of ΔF0 [$F(1,15) = 1.11$, $p = 0.308$], and the average value of d' (averaged across all ΔF0 values for each subject) was not significantly different from zero [one-sample t-test; $t(15) = 2.26$, $p =$

0.074]. In the across- sequence task (right panel), a significant main effect of ΔF0 was observed [$F(1,15) = 8.88$, $p = 0.009$], with a significant linear trend [$p = 0.05$], but the slope was positive, i.e., opposite to what would be expected based on the effects of streaming, and the overall level of performance was low (but better than chance on average). We have no clear explanation for why a positive slope emerged here.

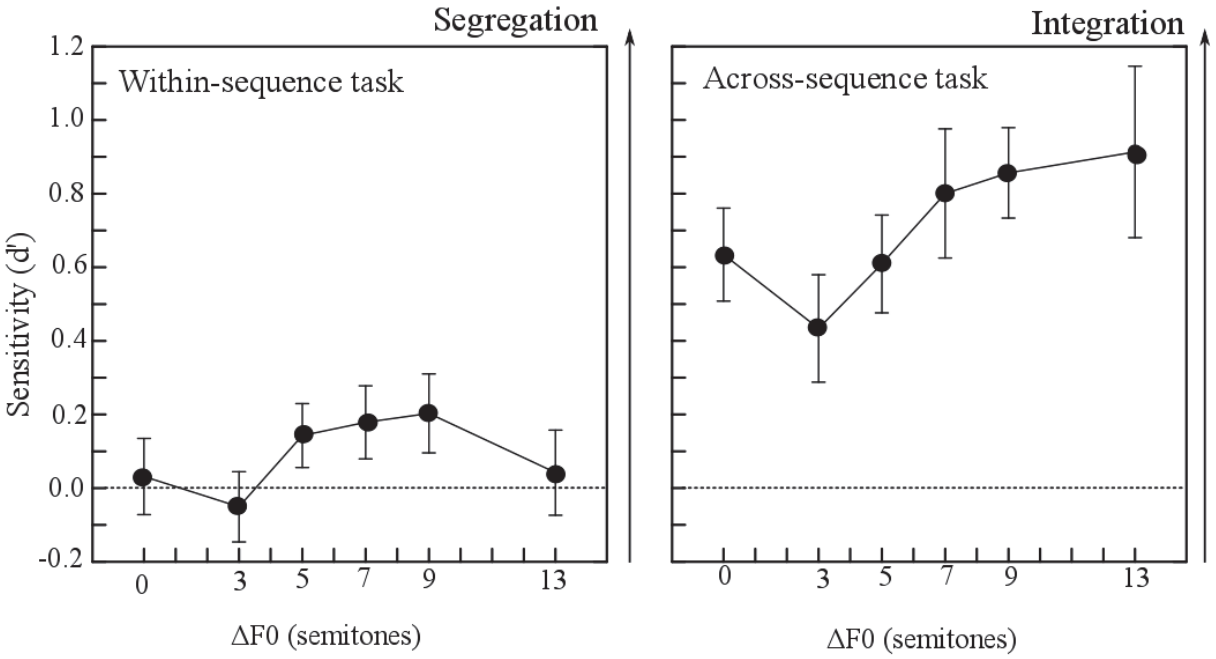


Fig. 6: Mean performance in terms of *d'*, averaged across the sixteen listeners in Experiment 3. The dotted line represents chance performance, and the error bars represent ±1 standard error of the mean.

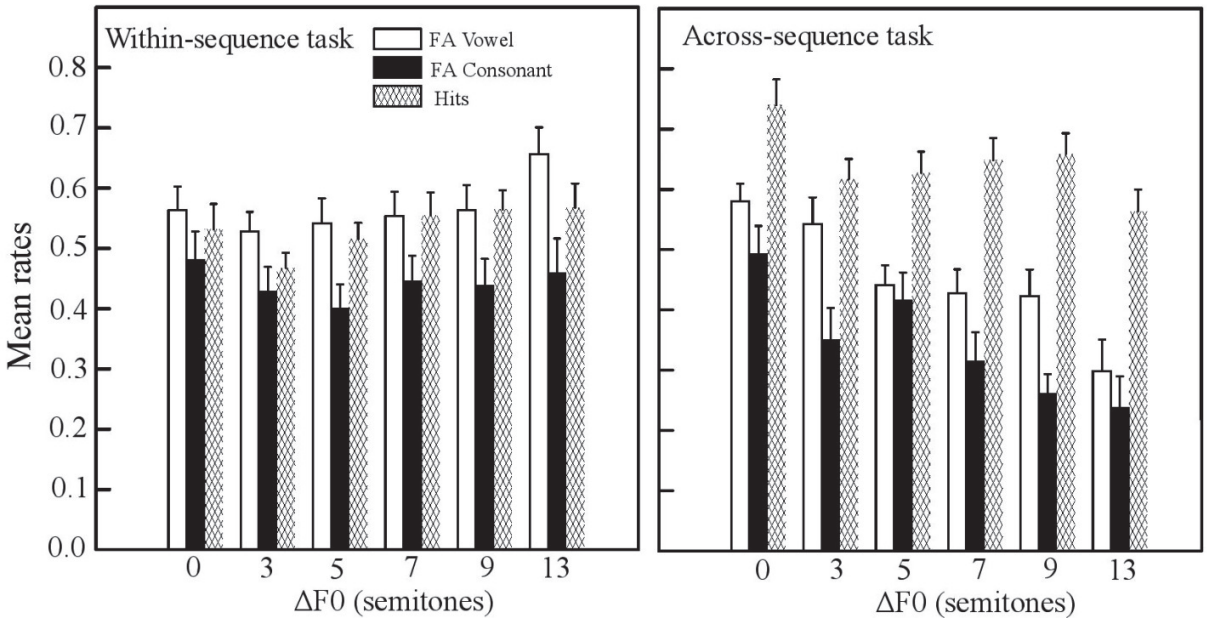
[FIG 6 ABOUT HERE]

Figure 7 shows the FA rates for both conditions, as a function of ΔF0. For both experiments, the FA rate was very high, reflecting the poor *d′* scores. A mixed-model ANOVA was performed on the values of *d′* for both tasks, with the order of the tasks (within- or across-sequence condition first) as a between-subjects factor, and FA type (vowel or consonant) and ΔF0 (0-13 semitones) as a within-subjects factors. The effect of FA type was significant in the within-sequence task [$F(1,14) = 17.2$, $p = 0.001$]. This result indicates that the vowels were responsible of the high FA rates, most likely because the vowels contain all the distinguishing acoustic cues. The effects of ΔF0 and the interactions were not significant. Considering now the across-sequence task, both the effect of FA type [$F(1,14) = 20.9$, $p < 0.001$] and the effect of ΔF0 [$F(1,14) = 40.3$, $p < 0.001$] were significant. The effect of the order of the task was not significant. The results of this experiment suggested that the vocalic formant transition by itself did not provide enough information to correctly identify the missing unvoiced part of the token.

Given the poor performance of listeners in this experiment when the unvoiced portions of the speech were removed, particularly in the within-sequence task, it seems that the results from Experiments 1 and 2 cannot easily be explained only in terms of speech information present in the voiced portions of the speech. Instead, a parsimonious account of all the data presented in this study is that listeners are able to perceptually segregate CV tokens based on differences in F0 that are present only in the voiced portions of the tokens. Spectrotemporal continuity, based on coarticulation, can contribute to the binding of the consonant and vowel portions of the CV tokens. Cole and Scott (1973) showed that the order of a sequence of CV tokens could not be accurately reported when the vowel transition was removed, indicating that the vowel transition facilitated the integration of the sequences. Similarly, by changing the formant transition and modifying the shape of the F0 contour of CVC syllables, Stachurski et al. (2015) found that both

487 cues affect the binding of consonants and vowels. Another contributing factor is likely to
488 involve perceived continuity induced by the vocal tract length (one single talker recorded the
489 whole syllable) (Tsuzaki et al., 2007). Regardless of the mechanism, the present study confirms
490 that such binding occurs and demonstrates that it can be used in perceptual stream segregation.

491



492

493 **Fig. 7:** Same as Fig. 4 with the results of Experiment 3.

494 [FIG. 7 ABOUT HERE]

495

496 **5. CONCLUSIONS**

497 This series of experiments tested whether differences in F0 could induce auditory stream
498 segregation between sequences of CV tokens, even though the unvoiced consonant part of the
499 CV contained no voiced information. The results can be summarized as follows:

500 • Experiment 1 showed that listeners could use F0 differences between syllables
501 containing an unvoiced fricative consonant and a voiced vowel (CV token) to form
502 perceptual streams. When the listeners' task encouraged segregation (voluntary
503 streaming), performance improved with increasing ΔF0; however, when the listeners'
504 task encouraged integration of the streams, increasing the ΔF0 from 0 to 9 semitones did
505 not lead to a significant decrement in performance, suggesting that obligatory streaming
506 did not occur. The relatively long (200-ms) onset-to-onset interval might have
507 contributed to this outcome.
508 • Experiment 2 investigated the possibility that listeners were basing their judgments on
509 either just the vowels or just the consonants, and increased the tested range of ΔF0 to 13
510 semitones. Again, evidence for voluntary streaming was found, suggesting that listeners
511 were indeed using both the consonant and vowel portions of the stimuli to perform the
512 task. The analysis of the FA rate found no evidence that listeners were basing their
513 judgments on the vowel only or on the consonant only. Even with the larger range of
514 ΔF0, effects of obligatory streaming failed to reach significance.
515 • Experiment 3 tested the possibility that listeners were able to extract the identity of the
516 consonant from just the voiced portion of the CV, by removing the unvoiced portion of
517 the stimuli. Performance was near chance in conditions requiring perceptual segregation
518 of the interleaved sequences, suggesting that the voiced portions of the tokens did not

519 carry sufficient information about the consonant to enable accurate performance in the
520 streaming task.

521 Overall, the results suggest that listeners are able to form sequential auditory streams of
522 alternating speech sounds based solely on F0 differences in the voiced portions of the speech.
523 The cues that enable the grouping of the unvoiced with the voiced portions of speech and their
524 segregation from competing sounds remain to be investigated further.

525

526 **ACKNOWLEDGMENTS**

527 This work was supported by NIH grant R01 DC007657 (AJO), Erasmus Mundus Auditory
528 Cognitive Neuroscience travel award 22130341 (MD), and LabEX CeLyA ANR-10-LABX-
529 0060/ANR-11-IDEX-0007 (ML, NG). We thank Matthew Winn for helpful discussions that
530 led to Experiment 3, as well as Brian Roberts and an anonymous reviewer for their
531 constructive comments to further improve the manuscript.

532

533 **REFERENCES**

534 Anstis, S., and Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones.
535 *J. Exp. Psychol.: Human Percept. Perform.*, *11*(3), 257–271.

536 Assmann, P. F., and Summerfield, Q. (1990). Modeling the perception of concurrent vowels:
537 Vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, *88*(2), 680-697.

538 Bird, J., and Darwin, C. J. (1998). Effects of a difference in fundamental frequency in separating

two sentences. In A. R. Palmer, A. Rees, Q. Summerfield, and R. Meddis (Eds.), *Psychophysical and Physiological Advances in Hearing* (pp. 263–269). London: Whurr, London.

Boersma, P., and Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot. International*, *5*(9–10), 341–345.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sounds*. (MIT Press, Ed.). Cambridge.

Brokx, J. P., and Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *J. Phon.*, *10*(1), 23–36.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.*, *109*(3), 1101–1109.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, *25*, 975–979.

Cole, R. A., and Scott, B. (1973). Perception of temporal order in speech: the role of vowel transitions. *Can. J. Psychol.*, *27*(4), 441–9.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J. Acoust. Soc. Am.*, *114*(5), 2913-2922.

Darwin, C. J., and Hukin, R. W. (2000). Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J. Acoust. Soc. Am.*, *107*(2), 970-977.

David, M., Grimault, N., and Lavandier, M. (2015). Sequential streaming, binaural cues and lateralization. *J. Acoust. Soc. Am.*, *138*(6), 3500–3512.

Ewert, S. (2013). AFC: A modular framework for running psychoacoustics experiments and computational perception models. In *International Conference in Acoustics AIA-DAGA (Merano, Italy)* (pp. 1326–1329).

Gaudrain, E., and Başkent, D. (2015). Factors limiting vocal-tract length discrimination in cochlear implant simulations. *J. Acoust. Soc. Am.*, *137*(3), 1298–308.

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*(1–2), 32–41.

Harris, K. (1958). Cues for the discrimination of american english fricatives in spoken syllables. *Language and Speech*, *1*(1), 1–7.

Hartmann, W. M., and Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Percept.*, *9*(2), 155–183.

Haywood, N. R., and Roberts, B. (2010). Build-up of the tendency to segregate auditory streams : Resetting effects evoked by a single deviant tone. *J. Acoust. Soc. Am.*, *128*(5), 3019–3031.

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am.*, *108*(3), 1252-1263.

Macmillan, N. A., and Creelman, Douglas, C. (2004). *Detection Theory: A User's Guide Neil A. Macmillan, C. Douglas Creelman*. (P. Press, Ed.) (2nd ed.). Psychology Press.

579 Micheyl, C., and Oxenham, A. J. (2010). Objective and subjective psychophysical measures of
580 auditory stream integration and segregation. *J. Assoc. Research in Otolaryngol.*, *11*(4), 709–
581 724.

582 Miller, G. A. (1957). The masking of speech. *Psychol. Bull.*, *44*(2), 105–129.

583 Moore, B. C. J., and Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta*
584 *Acust. United Ac.*, *88*(3), 320–333.

585 Moore, B. C. J., and Gockel, H. (2012). Properties of auditory stream formation. *Phil. Trans. R.*
586 *Soc. B.*, *367*, 919–931.

587 Repp, B. H. (1981). Two strategies in fricative discrimination. *Percept. & Psychophys.*, *30*(3),
588 217–227.

589 Roberts, B., Glasberg, B. R., and Moore, B. C. J. (2002). Primitive stream segregation of tone
590 sequences without differences in fundamental frequency or passband. *J. Acoust. Soc. Am.*,
591 *112*(5), 2074–2085.

592 Stachurski, M., Summers, R. J., and Roberts, B. (2015). The verbal transformation effect and the
593 perceptual organization of speech: Influence of formant transitions and F0-contour
594 continuity. *Hear. Res.*, *323*, 22–31.

595 Tsuzaki, M., Takeshima, C., Irino, T., and Patterson, R. D. (2007). Auditory Stream Segregation
596 Based on Speaker Size , and Identification of Size-Modulated Vowel. In B. Kollmeier, G.
597 M. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, & J. Verhey
598 (Eds.), *Hearing - From Sensory Processing to perception* (pp. 285–294). Berlin: Springer-
599 Verlag Berlin Heidelberg.

van Noorden, L. (1975). *Temporal Coherence in the Perception of Tone Sequences*. Institute for Perception Research. Eindhoven.

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*(2), 938–945.

Vliegen, J., and Oxenham, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *J. Acoust. Soc. Am.*, *105*(1), 339–346.

Wagner, A., Ernestus, M., and Cutler, A. (2006). Formant transitions in fricative identification: the role of native fricative inventory. *J. Acoust. Soc. Am.*, *120*(4), 2267–2277.

Warren,R. M. (1961). Illusory changes of distinct speech upon repetition-the verbal transformation effect. *British Journal of Psychology*, *52*(3), 249–258.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Percept. & Psychophys.*, *35*(1), 49–64.

Yumoto, E. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *J. Acoust. Soc. Am.*, *71*(6), 1544-1550.