Discrimination and streaming of speech sounds based on differences in interaural and spectral cues

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

Discrimination and streaming of speech sounds based on differences in interaural and spectral cues

Marion David[a)]
Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455, USA

Mathieu Lavandier
Univ Lyon, ENTPE, Laboratoire Génie Civil et bâtiment, Rue Maurice Audin, 69518 Vaulx-en-Velin Cedex, France

Nicolas Grimault
Centre de Recherche en Neurosciences de Lyon, Université Lyon 1, Cognition Auditive et Psychoacoustique, Avenue Tony Garnier, 69366 Lyon Cedex 07, France

Andrew J. Oxenham
Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455, USA

(Received 9 March 2017; revised 1 September 2017; accepted 7 September 2017; published online 27 September 2017)

Differences in spatial cues, including interaural time differences (ITDs), interaural level differences (ILDs) and spectral cues, can lead to stream segregation of alternating noise bursts. It is unknown how effective such cues are for streaming sounds with realistic spectro-temporal variations. In particular, it is not known whether the high-frequency spectral cues associated with elevation remain sufficiently robust under such conditions. To answer these questions, sequences of consonant-vowel tokens were generated and filtered by non-individualized head-related transfer functions to simulate the cues associated with different positions in the horizontal and median planes. A discrimination task showed that listeners could discriminate changes in interaural cues both when the stimulus remained constant and when it varied between presentations. However, discrimination of changes in spectral cues was much poorer in the presence of stimulus variability. A streaming task, based on the detection of repeated syllables in the presence of interfering syllables, revealed that listeners can use both interaural and spectral cues to segregate alternating syllable sequences, despite the large spectro-temporal differences between stimuli. However, only the full complement of spatial cues (ILDs, ITDs, and spectral cues) resulted in obligatory streaming in a task that encouraged listeners to integrate the tokens into a single stream.
 [http://dx.doi.org/10.1121/1.5003809]

[VB] Pages: 1674–1685

I. INTRODUCTION

Understanding speech in complex auditory backgrounds relies on our ability to perceptually organize competing sound sources into streams. In the case of speech, the sounds emanating from a target speaker must be grouped together (integration) and separated from the competing background (segregation) to be intelligible (Bregman, 1990). In an early study, Cherry (1953) demonstrated that spatial separation between a target speaker and a masker can improve speech recognition. Cherry used dichotic presentation, with the target presented to one ear and the masker presented to the other. In real auditory environments, localization in both the median and horizontal planes is achieved via more subtle cues, such as interaural time and level differences (ITDs and ILDs, respectively) and monaural spectral differences (Blauert, 1997; Wightman and Kistler, 1992). These cues can be characterized via the head-related transfer function (HRTF; e.g., Gardner and Martin, 1995).

Many studies have investigated streaming using ITDs and ILDs (Hartmann and Johnson, 1991; Darwin and Hukin, 1999; Gockel *et al.*, 1999; Oxenham, 2000; Roberts *et al.*, 2002; Sach and Bailey, 2004; Kidd *et al.*, 2005; Stainsby *et al.*, 2011; Füllgrabe and Moore, 2012). Fewer studies have investigated the effect of spectral cues produced by simulating sounds from different locations. However, those that have studied the effects of spectral spatial cues, independent of binaural cues, have found that alternating sequences of broadband noise bursts can be perceptually segregated based on small spectral differences between the stimuli (Middlebrooks and Onsan, 2012). Stream segregation based on these spectral cues can also be obligatory (David *et al.*, 2014; David *et al.*, 2015), in that segregation occurs even in situations where listeners are instructed to integrate the sequences into a single stream; for a discussion of voluntary and obligatory streaming, see Micheyl and Oxenham (2010).

Although subtle spectral cues may be sufficient to segregate spectrally uniform noise bursts, it is not clear if this finding generalizes to more realistic stimuli, such as speech. First, the spectral variations in speech might make the spectral cues from spatial location less reliable. Second, the

a)Electronic mail: david602@umn.edu

voiced portions of speech contain primarily low-frequency information, which will be less affected by the high-frequency spectral cues associated with spatial differences.

In the present study, speech sounds were used, which consisted of both unvoiced (fricative consonant) and voiced (vowel) parts. These consonant-vowel (CV) tokens were naturally uttered and randomly concatenated to form interleaved sequences. David *et al*. (2017) used the same stimuli to show that differences in fundamental frequency (F0), which affected primarily the lower-frequency voiced part of the stimulus, could induce streaming of the entire CV. In order to avoid a potentially confounding effect of F0 differences in the present study, all the stimuli had the same F0, while maintaining the natural variations in the spectral and temporal envelopes of speech. One question posed by the present study is whether the spectral cues that primarily affect the higher-frequency portions of the stimulus can also lead to streaming of the entire CV. Another question was the extent to which binaural cues in the horizontal plane contribute to stream segregation, over and above the monaural spectral cues that are also available in the horizontal plane (David *et al*., 2014). The experiments related to these questions were preceded by a discrimination task to ensure that listeners could perceive the differences induced by imposing different spatial or spectral cues on the stimuli. Depending on the cues available, these differences could be differences in spectrum (coloration) and/or perceived position.

II. EXPERIMENT 1: DISCRIMINATION TASK

A. Rationale

The aim of the discrimination task was to assess the extent to which listeners can perceive a difference in spatial or spectral cues between successive speech tokens, with and without between-token variability. In the horizontal plane, all the spatial cues (spectral differences, ILD and ITD) were available for the listener to discriminate the stimuli. Neither ILD nor ITD would be substantially affected by variability in the spectra of the tokens, so we predicted that listeners' discrimination performance should not be substantially affected. However, changes in source location within the median plane produce only spectral differences, which are more likely to be susceptible to interference by spectral variability between the tokens themselves. We used non-individualized HRTFs to produce changes in spectral cues that are representative of those elicited by stimuli presented at different elevations.

B. Method

1. Stimuli

The stimuli used were a subset of those used by David *et al*. (2017). The naturally uttered CV tokens (male voice) consisted of four different fricative consonants ([f], [s], [th] and [sh]) combined with nine different vowels ([æ], [e], [iː], [I], [ə], [ɛ], [ʌ], [ɑ] and [uː]). The stimuli were truncated to 160 ms by shortening both the fricative consonant and the vowel, so that each portion was approximately equal in length. The truncated segment was then gated on and off with 10-ms raised-cosine ramps. The F0 of the voiced portions was flattened to 110 Hz using the software Praat (Boersma and Weenink, 2017) and then the stimuli were 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. This process equalized the F0, while preserving the natural spectral- and temporal-envelope variations of the speech stimuli.

The stimuli were filtered with non-individualized HRTFs (Gardner and Martin, 1995) to simulate different positions in the horizontal and median planes. It is worth noting that the spectral cues associated with elevation might not have been necessarily attributed to clear perceived positions by the listeners due to the use of non-individualized HRTFs. Nevertheless, the spectral differences introduced by these HRTFs should be representative of those experienced by normal-hearing listeners.

The excitation patterns (Glasberg and Moore, 1990) of three processed tokens with the same vowel but different consonants ([sha], [fa] and [tha]) simulated at 0° azimuth and 0° elevation are presented in the left panel of Fig. 1. The right panel of Fig. 1 illustrates the mean excitation patterns of the spectrum, averaged across all tokens used in the study, simulated at three different positions in the median plane (0°, 30°, and 70°). The spectra in the left panel illustrate the large high-frequency variability from token to token, even when they share the same vowel and are presented with the same fixed F0. Indeed, comparing the left and right panels of Fig. 1, the spectral differences from token to token are often

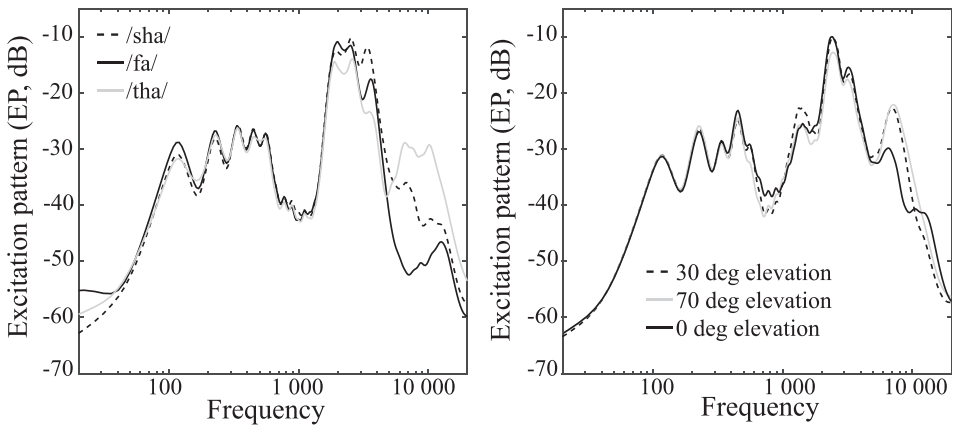

FIG. 1. Excitation patterns produced by different combinations of tokens and simulated spatial positions. The left panel shows the excitation patterns of three different tokens used in this study, simulated at 0° azimuth, 0° elevation. The dotted line corresponds to [sha], and the solid black and grey lines correspond to [fa] and [tha], respectively. The right panel shows mean excitation patterns of all the tokens used in this study, simulated at three different positions in the median plane. The black and grey solid lines correspond to 0° and 70°, respectively, and the dotted line corresponds to 30° elevation.

larger than the spectral differences induced by a difference in simulated position in the median plane.

2. Listeners

Sixteen listeners participated in the experiment (12 females, 4 males, aged from 18 to 28 years, median = 21). All of them were native speakers of American English, had normal hearing (i.e., pure-tone audiometric thresholds better than 20 dB hearing level (HL) at octave frequencies between 250 and 8000 Hz), and were paid for their participation. All listeners provided written informed consent and the protocol was approved by the Institutional Review Board of the University of Minnesota.

3. Procedure

A three-interval forced-choice procedure was used in which two stimuli were presented from a simulated location directly ahead (0° azimuth and elevation) and one stimulus was presented at a different simulated location in either the horizontal or median plane. The order of the three stimuli was selected at random on each trial and the stimuli were separated by 500-ms inter-stimulus intervals. The task involved indicating which of the three stimuli came from a different location. Six angles were tested in both planes: ±5°, ±10°, and ±30° in the horizontal plane, and ±10°, ±30°, +50°, and +70° in the median plane. In the constant-token condition, one speech token was selected at random on each trial and the same speech token was presented in all three intervals. In the different-token condition, three different speech tokens were selected at random (without replacement) on each trial and presented in the three intervals. Thus, in the constant-token condition, any change in the stimulus signified a change in simulated location, whereas in the different-token condition each interval involved spectral changes. Correct-answer feedback was provided after each trial.

The listeners completed two sessions of two hours each. Each session contained two separate blocks, one with constant tokens and one with different tokens. One session was used to test all conditions in the horizontal plane, and the other session was used to test all conditions in the median plane. The orders of the two sessions (horizontal/median) and two blocks within each session (same/different) were counterbalanced across the 16 subjects. For the constant-token conditions, four repetitions of each position and each token were presented, so that each listener completed 864 trials (4 repetitions with 6 angles and 36 tokens) in total. Listeners completed the same number of trials (864) for the different-token conditions, but the tokens were selected at random on each trial. Both sessions took place in a sound-attenuating booth. The stimulus presentation and response collection were controlled using the AFC software package (Ewert, 2013) under MATLAB (Mathworks, Natick, MA). The stimuli were converted to analog signals using a Lynx22 (Lynx Studio Technology, Costa Mesa, CA) 24-bit soundcard at a sampling rate of 44 100 Hz and were presented at 65 dB sound pressure level (SPL) via HD 650 headphones (Sennheiser, Old Lyme, CT).

C. Results

The proportion of correct responses was transformed into rationalized arcsine units (RAU) (Studebaker, 1985) to make them more suitable for parametric statistical analyses. The results, averaged across listeners, are shown in Fig. 2. The dashed line represents chance level and the black and grey circles represent the results from the constant- and different-token conditions, respectively.

The results in the horizontal plane are shown in the left panel of Fig. 2. A three-way repeated-measures analysis of variance (ANOVA) was performed with the RAU-transformed percent-correct values as the dependent variable and the condition (constant or different tokens), absolute angle (5°, 10°, and 30°), and hemisphere (negative/left or positive/right) as within-subjects factors. There were significant main effects of absolute angle [$F(2,60) = 271.0$, $p < 0.001$] and hemisphere [$F(2,60) = 4.92$, $p = 0.03$], but no effect of condition [$F(1,30) = 1.20$, $p = 0.28$]. The two-way interaction between absolute angle and hemisphere was significant [$F(2,60) = 6.41$, $p = 0.003$]. No other interactions were significant ($p > 0.26$ in all cases). These outcomes reflect the improvement in performance with increasing absolute angle and the slight asymmetry between the results from the left and right hemispheres, but no significant difference in performance between the constant- and different-token conditions.

The results from the median plane are shown in the right panel of Fig. 2. A two-way repeated-measures ANOVA was

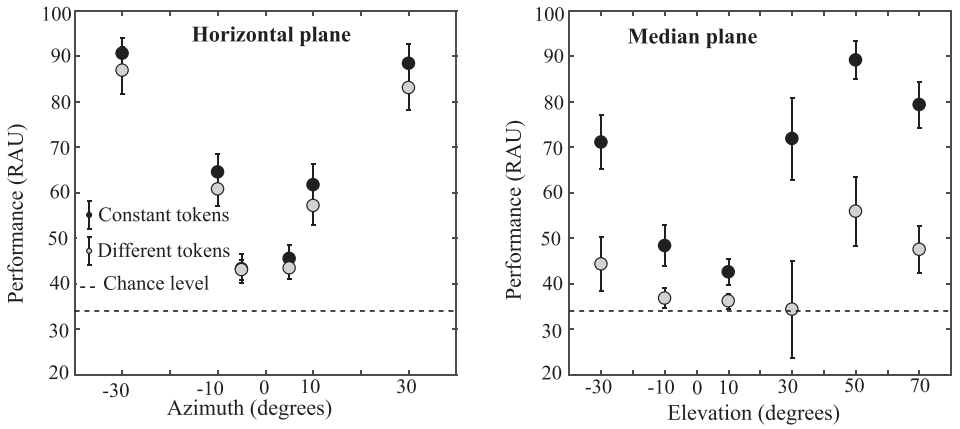

FIG. 2. Mean results from the discrimination task. Proportion correct, transformed into rationalized arcsine units (RAU), is shown as a function of angle. The left and right panels display the results for the horizontal and median planes, respectively. Black symbols represent results from the constant-token conditions, and grey symbols represent results from the different-token conditions. The dashed line represents chance level. Error bars represent ±1 standard error of the mean.

performed with the RAU-transformed percent-correct values as the dependent variable and the condition (constant or different tokens) and angle (–30°, −10°, +10°, +30°, +50°, and +70°) as within-subjects factors. Both main effects were highly significant [Condition: $F(1,15) = 77.7$, $p < 0.001$; Angle: $F(1,15) = 118$, $p < 0.001$], as was their interaction [$F(1,15) = 59.8$, $p < 0.001$]. Listeners performed significantly better when the stimuli did not vary from token to token. One-sample *t*-tests revealed that only performance for the stimuli simulated at +30° was not significantly above chance (33% or 34.21 RAU) in the different-token condition. For all other angles in this condition, mean performance was slightly but significantly above chance ($p < 0.008$ in all cases), even when accounting for multiple (6) comparisons using a Bonferroni correction ($\alpha = 0.05/6 = 0.0083$). Even though performance was generally quite poor in the different-token conditions, with mean scores between 37 and 56 RAU, there was some evidence that discrimination was still possible in the median plane.

D. Discussion

In the horizontal plane, performance improved as the difference in simulated position increased between the reference (0° azimuth) and the target. Regardless of whether the tokens were constant or different within each trial, a separation of 5° was sufficient to enable their discrimination. This level of performance is expected, given that the minimum audible angle (MAA) for broadband sounds is typically around 2.5° (Perrott and Pacheco, 1989), and that the primary cues for localization in the horizontal plane are ITD and ILD, which are not affected by whether the tokens are different or the same.

In the median plane, overall performance was poorer and the difference between the constant- and different-token conditions was greater. The poorer overall performance is expected, given that minimum audible angles in the median plane are generally higher, at around 4° to 9° (Perrott and Saberi, 1990). In addition, non-individualized HRTFs give a good approximation of the binaural cues (ILD and ITD) but are less accurate for the spectral cues produced by the pinnae, which vary substantially between individuals. Thus, because non-individualized HRTFs were used, differences in source elevation were potentially only perceived as a change in spectral coloration rather than a shift in the perceived location of the source. The HRTFs may also explain why performance was generally better at the 50° separation than at the 70° separation. A comparison of the differences in excitation patterns (Glasberg and Moore, 1990) between 0° and 50° and between 0° and 70° shows that the overall differences were greater for the smaller angle, with a mean absolute level difference of 2.85 dB for the smaller angle difference compared with an absolute level difference of about 2.31 dB for the larger angle difference (see Fig. 3).

The large detrimental effect of varying the tokens between intervals can be explained by the fact that the spectral differences between tokens interfered with the spectral differences imposed by the HRTFs, which were the only discrimination cue available for conditions in the median plane. Nevertheless, some discrimination from the reference remained possible at most tested elevations, leaving open the possibility that these cues could be used for auditory stream segregation, even in the presence of spectral variability of the tokens. This result is broadly consistent with the findings of Rakerd *et al.* (1999), who found that listeners were able to identify sounds with different spectral shapes when they all originated from the same location in space but were less able to perform the task when the location of the sounds in the median plane was randomly varied across presentations. Nevertheless, using sound sources in real space (rather than simulated HRTFs), they found that sound localization was possible even when listeners were not able to identify the sounds. The following experiment tested whether streaming was still possible with non-individualized HRTFs and with stimuli that varied in spectral shape between tokens.

III. EXPERIMENT 2: STREAM SEGREGATION USING BINAURAL AND SPECTRAL CUES

A. Rationale

Experiment 1 showed that listeners were able to detect changes in binaural and spectral cues in the horizontal plane, and changes in spectral cues in the median plane in some conditions, even in the presence of variability between tokens. The aim of experiment 2 was to determine whether listeners are able to use these changes to perceptually segregate alternating sequences of the CV tokens into streams.

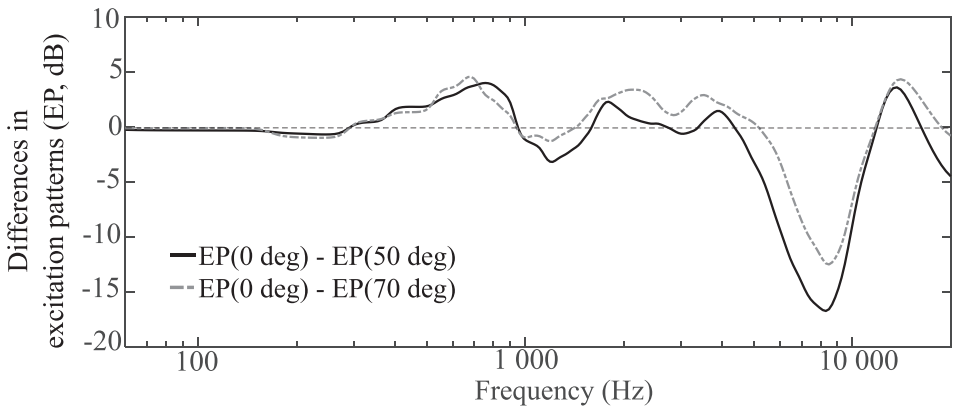

FIG. 3. Differences in excitation patterns following filtering by the HRTF between sounds incident from 0° and 50° (black curve), and from 0° and 70° (grey curve). The larger absolute difference between 0° and 50° may explain why average listener performance was better when the B tokens were presented from 50° than when the tokens were presented from 70°.

B. Method

1. Stimuli

The stimulus tokens used in experiment 2 were the same as those in experiment 1. Listeners were presented with two interleaved sequences of tokens alternating in simulated position (ABAB… sequences). Each token lasted 160 ms and was separated from the following token by a silent interval of 40 ms (interval between the end of a B and the beginning of the following A). The inter-onset time of 200 ms between successive tokens was short enough to observe some obligatory stream segregation (van Noorden, 1975). To encourage listeners to attend to the entire interleaved sequence on each trial, the length of each sequence varied randomly between 16 and 28 tokens (i.e., between 8 and 14 pairs). The speech tokens for the entire interleaved sequence were selected at random, without replacement, from the initial set of 36 tokens. The simulated position of the A sequence was constant at 0° azimuth and 0° elevation. The simulated position of the B tokens was selected from one of the following locations: 0°, 5°, 10° and 30° to the right in the horizontal plane (0° in elevation), or 0°, 10°, 50° and 70° elevation in the median plane (0° in azimuth).

2. Procedure

The listeners participated in two types of trials. In the within-sequence trials, listeners were asked to detect a consecutive repetition of one of the CV tokens that occurred within the B sequence (i.e., those tokens not emanating from the 0° location). In the across-sequence task, listeners were asked to detect a CV repetition that occurred between an A token and the following B token. Figure 4 provides a schematic diagram of the stimuli and tasks. In half of the trials, none of the tokens was repeated. In the other half, selected at random, a repetition of one CV token was introduced, as described above. Before each interleaved sequence, a short cueing sequence was presented, consisting of four tokens with an inter-stimulus interval of 240 ms. The four tokens in the cueing sequence were presented from the location of the B tokens in the main interleaved sequence. A gap of 1 s separated the end of the cueing sequence from the beginning of the main interleaved sequence.

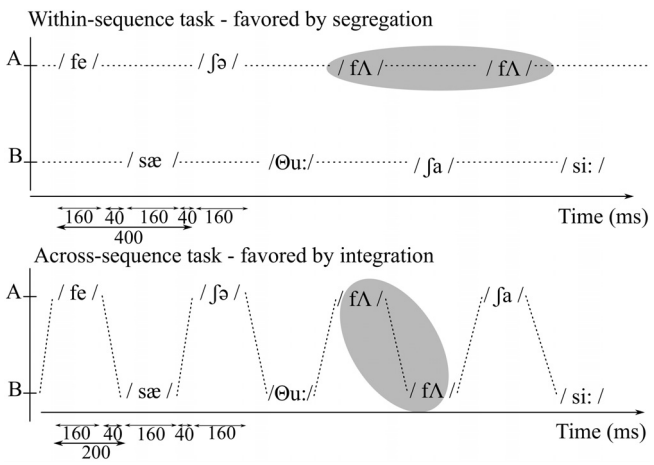

FIG. 4. Schematic diagram of the tokens in the within- and across-sequence tasks (upper and lower panels, respectively). The circled syllables correspond to a repeated CV token. In half of the trials, the sequences consisted of only different stimuli (not shown) and in the other half, a repetition was introduced. In the within-sequence task, performance was expected to improve with increasing perceived differences between the A and B tokens, whereas in the across-sequence task, performance was expected to improve with decreasing perceived differences.

If present, the repetition was always introduced immediately before the final pair of tokens to allow time for the build-up of segregation (if any) to occur (Anstis and Saida, 1985; Haywood and Roberts, 2010). As the length of the sequence was randomized between 8 and 14 pairs of tokens, the position of the repetition varied randomly between the 7th and the 13th pair. For good performance in the within-sequence task, listeners should perceptually segregate the A and B sequences into separate streams, and selectively attend to the stream containing the B tokens; thus, it provided a measure of voluntary stream segregation. Conversely, for good performance in the across-sequence task, listeners should perceptually integrate the A and B sequences into a single stream, so that the repetition is heard within this stream, making it a measure of obligatory stream segregation (Micheyl and Oxenham, 2010; van Noorden, 1975).

In both tasks, the listeners had to indicate whether or not the interleaved sequence contained a repetition of a CV token. Feedback was provided after each trial. The hit rate (H) was defined as the proportion of correctly detected repetitions and the false alarm rate (FA) corresponded to the proportion of trials with no repetition in which a repetition was reported. Listeners' sensitivity to the repetition (d') was estimated by subtracting the z-transform (i.e., the inverse cumulative normal distribution function) of FA from the z-transform of H. A correction was applied when H was 100% or FA was 0% using $1–1/(2N)$ and $1/(2N)$, respectively, where N is the total number of trials (Macmillan and Creelman, 2004).

Prior to undertaking the main experiment, the participants in this experiment completed a 2-h pilot session to test the feasibility of the approach. During this pilot session, three extreme conditions in the horizontal plane and two in the median plane were tested. The first condition tested in the horizontal plane was a single-sequence (i.e., B-tokens only) condition. This condition provided a baseline comparable to a condition with perfect segregation of the A and B sequences. In this condition, the listener had to detect whether a repeat was introduced in the sequence; as only one sequence was presented, the inter-stimulus interval (ISI) was 240 ms. The second condition consisted of one sequence (A-tokens) simulated at 90° to the left and the other sequence (B-tokens) simulated at 90° to the right. The listeners had to detect whether or not a repeat was present in the sequence coming from the right. In this condition, the sequences should be mostly segregated and thus the tokens within each stream would be heard with an ISI of 240 ms. No difference between the sequences was introduced in the third condition (i.e., A- and B-sequences were both simulated at the same position 0° elevation and 0° azimuth). In this condition segregation should not have been possible, so sequences were mostly fused and the tokens were heard with an ISI of 40 ms. The

listeners were asked to detect a repeat across the two integrated sequences. This condition represents an extreme case of integration.

In the median plane, two conditions were tested during the pilot session. The first consisted of one sequence simulated at 0° elevation and the other sequence simulated at 90° elevation. The listeners had to focus on the upper sequence to detect whether or not a repeat was presented. Finally, the condition without a difference between the sequences was tested with both sequences simulated at 0° elevation and 0° azimuth. The stimuli in this condition were identical to those used in the final condition in the horizontal plane.

Eight blocks were completed per plane, with each block containing six repetitions of each of the three or two extreme conditions, for a total of 288 and 192 trials for the horizontal and median planes, respectively, half with a repeat and half without. All the participants performed above chance in all the pilot conditions. Since the sensitivity d' was above 1 in all the conditions, we concluded that the tasks were feasible in both planes.

After the pilot session, listeners completed two 2-h sessions; each session was devoted to one plane (horizontal or median) and the order of the four tasks (horizontal across, horizontal within, median across, and median within) was counterbalanced between listeners. Fourteen blocks were completed for each task. For the horizontal plane, each block contained three repetitions of the five conditions (one single sequence of B-tokens, 0°, 5°, 10° and 30° azimuth at 0° elevation) for a total of 420 trials (210 with repeat and 210 without) per task. For the single sequence, the angle from where the sequence was simulated was randomly chosen; the elevation was fixed to 0° in all the horizontal tasks and the azimuth was fixed to 0° in all the median tasks. For the median plane, each block contained three repetitions of the four conditions (0°, 10°, 50° and 70° elevation and 0° azimuth) for a total of 336 trials (168 with repeat and 168 without) per task. The experimental setup was the same as for experiment 1.

3. Listeners

Ten native speakers of American English participated in this experiment (6 females, 4 males, aged from 18 to 28 year, median = 19). All of them had normal hearing (i.e., pure-tone thresholds better than 20 dB HL at octave frequencies between 250 and 8000 Hz), and were paid for their participation. None of them had previously participated in the discrimination task (experiment 1) but all of them participated in both the pilot and test sessions. All listeners provided written informed consent and the protocol was approved by the Institutional Review Board of the University of Minnesota.

C. Results

The results, averaged across listeners, are shown in Fig. 5 as a function of the simulated spatial separation between the A and B token sequences (azimuth in the horizontal plane and elevation in the median plane). The upper panels correspond to the results in the horizontal plane and the lower panels correspond to the results in the median plane. The left and right columns represent results for the within- and across-sequence tasks, respectively.[1]

Performance in the within-sequence task improved with increasing the simulated separation angle. In the horizontal plane (Fig. 5, upper-left panel), a repeated-measures ANOVA revealed a significant main effect of simulated spatial separation [$F(3,27) = 6.02$, $p = 0.003$]. In the median plane (Fig. 5, lower-left panel), the main effect of simulated position was also significant [$F(3,27) = 3.32$, $p = 0.035$].

Performance in the across-sequence task seemed to decrease when increasing the simulated spatial separation between the sequences. In the horizontal plane (Fig. 5, upper-right panel), a repeated-measures ANOVA revealed a main effect of simulated spatial separation [$F(3,27) = 13.95$, $p < 0.001$]. In the median plane (Fig. 5, lower-right panel), however, no significant main effect of simulated position was observed [$F(3,27) = 2.04$, $p = 0.131$], although a trend in the same direction was apparent.

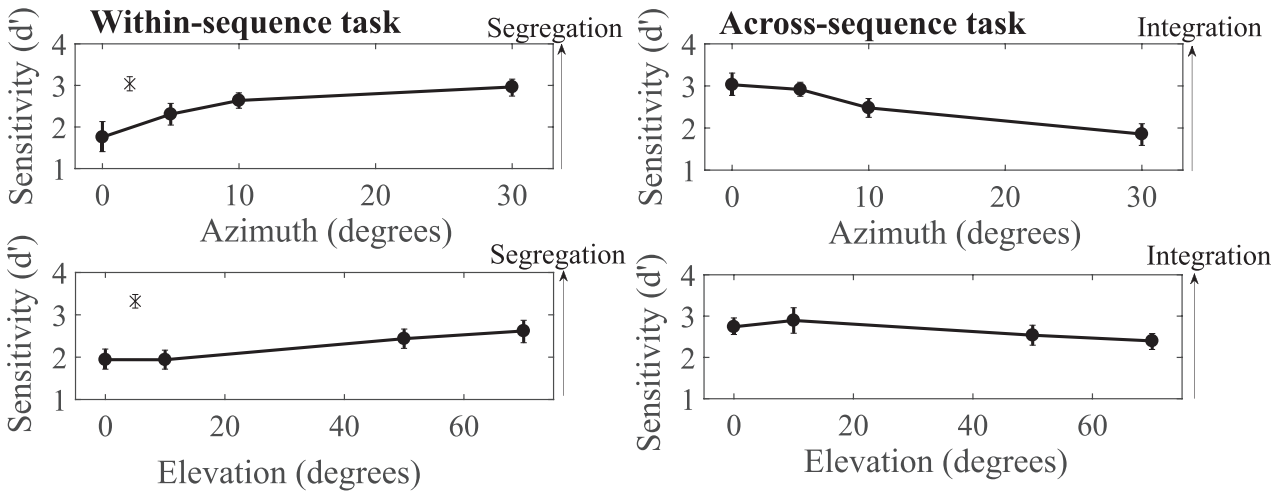

FIG. 5. Mean performance in terms of d' for the within-sequence (left column) and across-sequence (right column) tasks in the horizontal plane (top panels) and median plane (bottom panels) in experiment 2. In the within-sequence task, a high d' indicates a greater tendency to segregate the sequence into two different streams. In the across-sequence task, a high d' indicates a greater ability to integrate the sequence into one single stream. Crosses in the left panels indicate estimated sensitivity in the single-sequence conditions (see footnote 1). Error bars correspond to ±1 standard error of the mean.

D. Subjective task

Eight of the ten listeners who participated in experiment 2 also performed a subjective task. Results from the objective task suggested that voluntary (and to some extent obligatory) segregation increased with increases in simulated spatial separation. The aim of the subjective task was to provide a more direct measure of perceived segregation with these stimuli. In cases of stream segregation, the listeners should hear two separate voices, whereas when the sequences integrate into a single stream they should hear a single voice. Thus, the same sequences as in experiment 2 were presented and listeners were asked at the end of each sequence to indicate whether they heard one voice or two voices (Micheyl and Oxenham, 2010).

Twelve blocks were completed for each plane (horizontal and median). Each block contained four repetitions of each simulated position (0°, 5°, 10° and 30° in the horizontal plane and 0°, 10°, 50°, 70° in the median plane), so that listeners had to judge 192 sequences in each plane. Figure 6 displays the results of the subjective experiment. The results of both the objective and subjective tasks were consistent. In both planes, integration decreased as the difference in simulated position increased, although again the results were less compelling in the median plane. In this plane, the spectral differences induced by the difference in simulated position are often smaller than the spectral variations from token to token. This might explain why only 55% of the sequences at the 0° separation in this plane were judged as being "one voice," whereas in the identical condition in the horizontal plane over 80% of sequences were judged as being "one voice." Nevertheless, one-way ANOVAs confirmed a significant effect of simulated separation on judgments for both the horizontal plane [$F(3,33) = 15.17$, $p < 0.001$] and the median plane [$F(3,33) = 10.84$, $p < 0.001$].

E. Discussion

So far, this study has shown that listeners can perceive the regularities in interaural and spectral differences induced by a difference in simulated spatial location, despite large spectral variability from token to token (experiment 1), and that these regularities can be extracted to form auditory streams (experiment 2). This outcome is particularly interesting in the case of spectral differences associated with simulated spatial differences in the median plane, as it extends the results of Middlebrooks and Onsan (2012), David *et al.* (2014), and David *et al.* (2015) by showing that monaural spectral cues can induce streaming even in the presence of natural spectral variability between tokens. Martin *et al.* (2012) showed only a slight spatial release from masking in the median plane using speech filtered by individualized head related transfer functions. However, they did not look at streaming *per se*, but at masking release, which may involve mechanisms of segregation of simultaneous, as well as sequential sounds.

One interesting aspect of the data is that segregation of the CVs appears possible based on differences in simulated location along the median plane, even though the spectral differences occurred mainly at high frequencies, in regions dominated primarily by the consonant portion of the CV (see Fig. 1). This outcome suggests that the consonant and vowel parts of the tokens were perceptually bound, in line with the earlier findings of David *et al.* (2017). In their study, changes in F0, which were limited to the voiced (vowel) portions of the stimulus, were sufficient to induce stream segregation of the entire CV. In the present case, the converse also appears to hold: streaming cues limited primarily to the consonant portion are sufficient to induce segregation for the entire CV. However, it is also possible that the participants were basing their judgments solely on the consonants and ignoring the vowels in order to perform well in the within-sequence task. To rule out this possibility, experiment 3 was run in an attempt to distinguish the separate contributions of the vowels and consonants to performance in the streaming task.

IV. EXPERIMENT 3: SEPARATE CONTRIBUTION OF VOWELS AND CONSONANTS IN THE STREAMING TASK

A. Rationale

The aim of this experiment was to determine whether listeners segregate the entire CV token in each sequence, or whether they perform the task instead by attending only to the consonant or only to the vowel portion of each token. To answer this question, the 50% of trials without repetitions of experiment 2 were replaced here by trials containing a repetition of either just the consonant or just the vowel. According to this paradigm, also used in David *et al.* (2017), high performance would be possible only if the listeners were able to attend to the correct sequence and perceive the

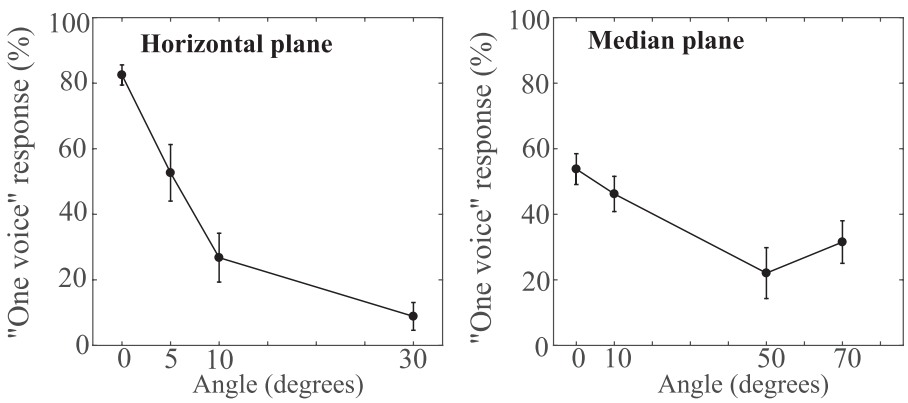

FIG. 6. Mean proportion of "one voice" responses for the subjective task, where the listeners had to indicate whether they heard one or two voices at the end of the presented sequences. The left and right panels represent the results for the horizontal and median planes, respectively. Error bars represent ±1 standard error of the mean.

repetition of the whole token. The CV tokens used in this experiment were the same as those used in experiment 2, and were made from 9 vowels and 4 consonants.

B. Method

1. Procedure

As in experiment 2, listeners were presented with interleaved sequences of CV tokens from a variety of different simulated locations preceded by a 4-token cueing sequence. In 50% of the trials, randomly selected, a repetition of a full token (consonant and vowel, "full repeat") was presented. In 25% of the trials a repetition of only the consonant was presented, and in the last 25% of trials a repetition of only the vowel was presented. The last two cases are referred to as "half-repeat." According to this paradigm, the H corresponds to the proportion of full repeats that were correctly reported and FA corresponds to the proportion of trials in which a repetition was reported when only a half-repeat was presented. Thus, it was possible to calculate separately the FA for the consonant-only and vowel-only repeats.

As in experiment 2, two interleaved sequences (A and B) simulated at different positions were presented. For the within-sequence task, the listeners were asked to attend to the sequence that started first, as presented in the cueing tokens. For the across-sequence task, in both planes, the listeners were instructed to attend to the whole interleaved sequence, regardless of the simulated positions of the tokens. In all tasks (horizontal/median plane within/across-sequence task), the listeners had to indicate whether or not the interleaved sequence contained a repetition of the full CV token.

The listeners completed two 2-h sessions. Each session was devoted to one plane (horizontal or median), and the order of the four tasks (horizontal within/across-sequence tasks, median within/across-sequence tasks) was counterbalanced between listeners. Fourteen blocks were completed per task. For the horizontal plane, each block contained three repetitions of the five conditions (one single sequence, 0°, 5°, 10° and 30° azimuth at 0° elevation) for a total of 420 trials (210 with full-repeat and 210 with half-repeat). Note that for the single-sequence condition, the position from which the sequence was simulated was randomly chosen on each trial. For the median plane, each block contained three repetitions of the four conditions (0°, 10°, 50°, and 70° elevation at 0° azimuth) for a total of 168 trials with a full repeat and 168 with a half repeat. The experimental setup was the same as for experiments 1 and 2.

2. Listeners

Eight native speakers of American English participated in the present experiment (4 females, 4 males, aged from 19 to 64 year, median = 23). They all had audiometric thresholds better than 20 dB HL at octave frequencies between 250 and 8000 Hz and were paid for their participation. One participant had previously participated in experiment 2. All listeners provided written informed consent and the protocol was approved by the Institutional Review Board of the University of Minnesota.

C. Results and discussion

The d' scores, averaged across listeners, are shown in Fig. 7 as a function of the simulated spatial separation between the A and B token sequences. The upper panels correspond to the results in the horizontal plane and the lower panels correspond to the results in the median plane. The left and right columns represent the within- and across-sequence tasks, respectively.[2]

Overall, performance was somewhat poorer in this experiment than in experiment 2, where the no-repeat trials did not have a repeat of a vowel or a consonant. Performance in the within-sequence task improved when increasing the simulated separation angle in both planes, in line with the predictions of voluntary stream segregation

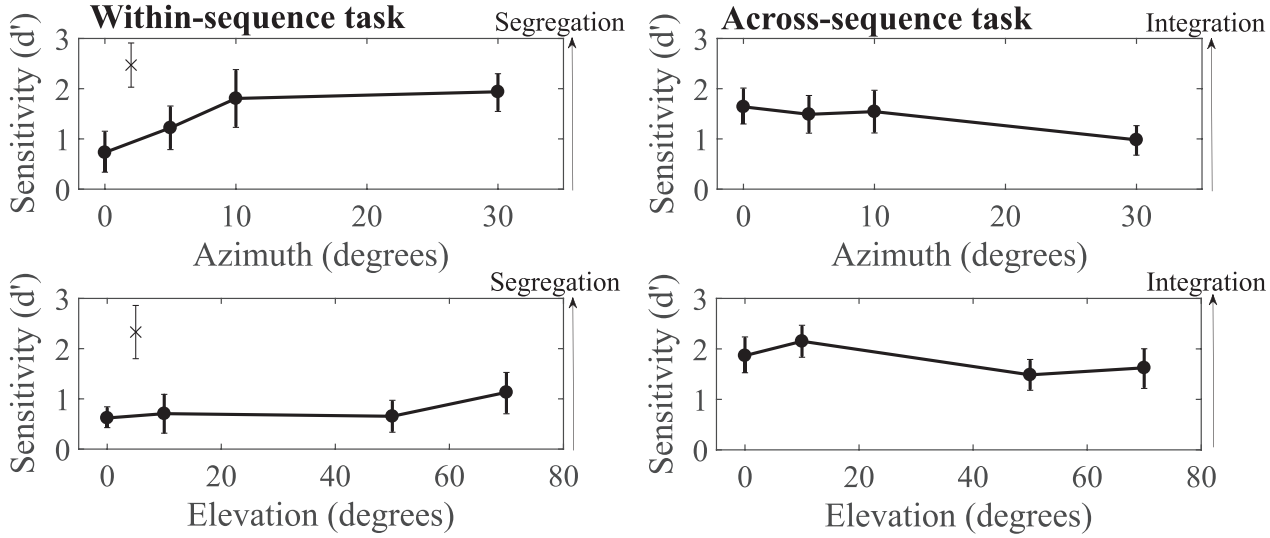

FIG. 7. Mean performance in terms of d' for the within- (left column) and across- (right column) sequence tasks in the horizontal plane (top panels) and median plane (bottom panels) in experiment 3. Crosses in the left panels indicate estimated sensitivity in the single-sequence conditions (see footnote 2). The error bars correspond to ±1 standard error of the mean.

based on spatial cues (Fig. 7, left panels). A repeated-measures ANOVA revealed a main effect of the simulated spatial separation in the horizontal plane [$F(3,21) = 12.1$, $p < 0.001$]. In the median plane, however, the trend toward an increase in performance with increasing the simulated spatial separation was not confirmed by the statistical analysis, which showed no main effect of spatial separation [$F(3,21) = 1.52$, $p = 0.24$].

Otherwise, the pattern of results appeared to be quite similar to that observed in experiment 2. To perform a direct comparison, repeated-measures ANOVAs were performed on the d' scores, combining data from the within-sequence experiment of both Experiments 2 and 3. A repeated-measures ANOVA with experiment as a between-subjects factor (ignoring the fact that one subject performed both experiments) revealed a main effect of experiment (in line with poorer performance in experiment 3) [$F(1,16) = 90.1$, $p < 0.001$], and a main effect of the simulated spatial separation [$F(3,48) = 15.0$, $p < 0.001$]. There was no significant interaction between simulated spatial separation and experiment [$F(3,48) = 0.16$, $p = 0.92$], consistent with the similar pattern of results across the two experiments. In the median plane, there was a main effect of experiment [$F(1,16) = 106.3$, $p < 0.001$] and a main effect of spatial separation [$F(3,48) = 3.92$, $p = 0.014$], but there was no significant interaction between spatial separation and experiment [$F(3,48) = 0.78$, $p = 0.51$], also in line with the similar pattern of results across the two experiments.

In the across-sequence task, as in the within-sequence task, performance was somewhat poorer overall than in experiment 2. A repeated-measures ANOVA revealed a main effect of the simulated spatial separation in both the horizontal and median planes [$F(3,21) = 6.53$, $p = 0.003$, $F(3,21) = 3.41$, $p = 0.036$, respectively].

Combining the data from the across-sequence conditions in experiments 2 and 3 in the horizontal plane, a repeated-measures ANOVA with experiment as a between-subjects factor showed a significant effect of the experiment [$F(1,16) = 129.0$, $p < 0.001$]. The main effect of the simulated spatial separation was significant [$F(3,48) = 18.1$, $p < 0.001$]. There was no significant interaction between the simulated spatial separation and experiment [$F(3,48) = 2.35$, $p = 0.084$]. In the median plane, there was a main effect of experiment [$F(1,16) = 186.3$, $p < 0.001$] and simulated position [$F(1,16) = 5.01$, $p = 0.004$]. Again, there was no significant interaction between simulated spatial separation and experiment [$F(3,48) = 0.39$, $p = 0.76$].

Overall, the lack of interactions between experiment and spatial separation confirms the impression that the pattern of results was similar in experiments 2 and 3. Therefore, it seems that the streaming effects observed in experiment 2 were not due to listeners attending only to the consonants or only to the vowels, but instead can be ascribed to the perceptual segregation and streaming of the entire CV. To test whether vowels or consonants were dominant in determining overall performance, an analysis of the FA patterns was conducted. The FA in response to a consonant-only or a vowel-only repeat is shown in Fig. 8, along with the H. The figure shows that neither the FA for the consonant-only nor the FA for the vowel-only trials dominated performance. Nonetheless, the FAs associated with the vowels seem slightly but consistently higher than the FA associated with the consonants in the within-sequence task in the median plane. Repeated-measures ANOVAs were performed separately for the data in the horizontal and median plane, with FA as the dependent variable and the factors FA type (consonant or vowel) and position (4 levels). Neither the main effects of FA type or simulated position nor their interactions were significant in the horizontal

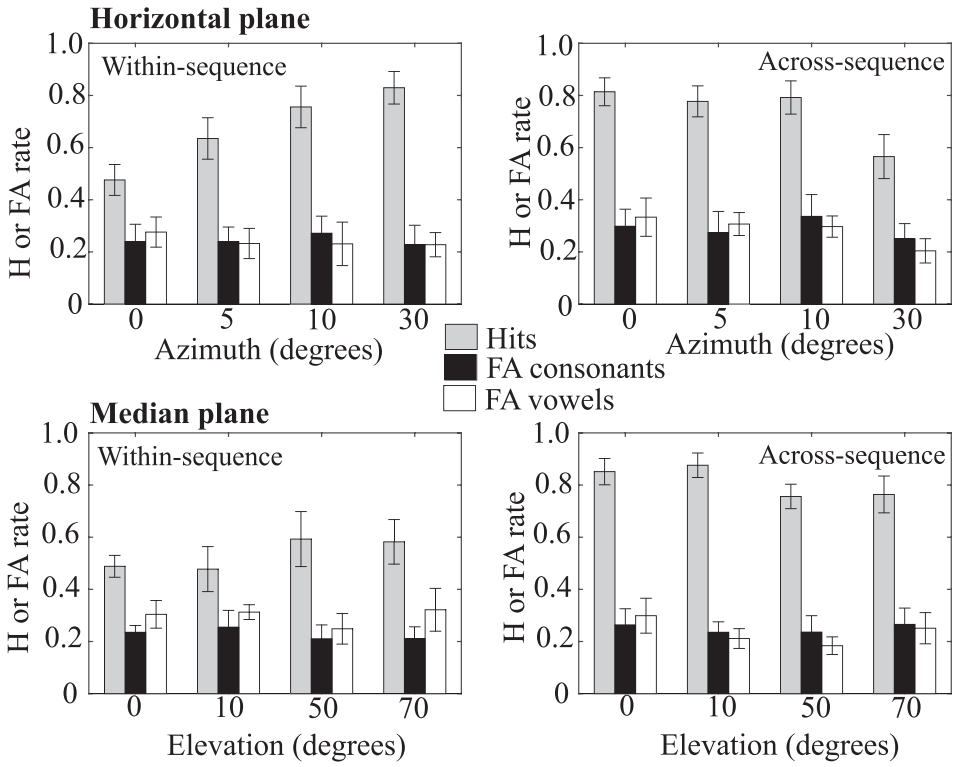

FIG. 8. Mean hit (H) and false-alarm (FA) rates for the within- and across-sequence tasks (left and right columns, respectively) for the horizontal (top panels) and median (bottom panels) planes in experiment 3. The error bars correspond to ± 1 standard error of the mean. The grey, black, and white bars correspond to the H, FA due to the consonants, and FA due to the vowels, respectively.

plane ($p > 0.29$ in all cases). In the median plane the main effect of FA type was significant in the within-sequence task [$F(1,7) = 7.34$, $p = 0.03$] showing that the FA were due more to vowels in that case. However, this effect was small, as shown in Fig. 8. The other condition (across-sequence task) did not show a significant effect of FA type, simulated position, or their interaction ($p > 0.52$ in all cases). These results suggest roughly equal contributions of the vowel and consonant portions of the stimuli to streaming in all conditions.

In summary, the results suggest that listeners made use of the entire CV, rather than just the vowel or consonant, in detecting the repeated token. This result is consistent with a previous study using CV tokens separated by a difference in F0 (David *et al.*, 2017).

V. EXPERIMENT 4: DETERMINING THE IMPORTANCE OF DIFFERENT SPATIAL CUES IN THE HORIZONTAL PLANE

A. Rationale

Experiments 2 and 3 showed that a difference in simulated positions in the horizontal plane can elicit stream segregation of CV tokens. In the horizontal plane, the potential cues for segregation are the spectral differences, the ILD and the ITD (e.g., Middlebrooks and Green, 1991) as well as the perceived position of the simulated source. Some studies have failed to show an effect of ITD on stream segregation, particularly in case of obligatory stream segregation of pure or complex tones (Füllgrabe and Moore, 2012; Stainsby *et al.*, 2011). However, monaural cues, as well as binaural cues, have been shown to have an influence on stream segregation of speech-shaped noises (David *et al.*, 2014; David *et al.*, 2015; Middlebrooks and Onsan, 2012). In the present experiment, the spectral and binaural cues were introduced progressively to assess the extent to which they were useful for streaming CV tokens.

B. Method

1. Stimuli and Procedure

Only the tokens simulated in the horizontal plane were considered in this experiment. The spectral differences, ILDs, and ITDs were introduced progressively to test their influence on streaming. Four conditions were tested. In the first condition, the stimuli did not contain interaural differences, but instead provided to both ears the same spectral cues associated with differences in simulated position, based on the spectrum measured at the left ear for the non-individualized HRTFs. This condition is referred to as "SPEC." In the second condition, the ILD was isolated by combining the magnitude of the lateral HRTFs with the phase of the 0° HRTFs before computing an inverse FFT. This manipulation set the ITDs to 0 while preserving the ILDs. This condition will be referred to as "ILD" (and still contained spectral cues). In the third condition the ITD was isolated by combining the phase of the lateral HRTFs with the magnitude of the 0° HRTFs and converting back to impulse response using inverse FFT (Culling and Mansell, 2013). This condition will be referred to as "ITD" (and still contained spectral cues). And finally, in the fourth condition the full complement of spatial cues was provided, including ITD, ILD, and spectrum, as in experiments 1, 2, and 3. This condition is referred to as "ALL."

The procedure and CV tokens were the same as in experiment 2, since experiment 3 showed that performance was not dominated by either the vowel or consonant alone. Two sequences of speech sounds alternated over time to form an interleaved sequence. As in experiment 2, one sequence (A) was always simulated as coming from the center (0°) and never contained any repeated tokens. The second sequence (B) was simulated as coming from a position to the left. Data were collected using simulated positions of 0°, 30°, and 90°. Listeners were asked to attend to the sequence coming from the left in the within-sequence task and to attend to the whole interleaved sequence in the across-sequence task. In the SPEC condition, the listeners were asked to focus on the sequence that started first in the within-sequence task. At the end of the interleaved sequence, they had to indicate whether or not they heard a repeated token.

The listeners completed one 2-h session. The order of the tasks (within- and across-sequence task) was counterbalanced between listeners. Eighteen blocks were completed per task, consisting of six repeats of the four conditions (SPEC, ILD, ITD, and ALL). For the within-sequence task, each block contained six repetitions of the four configurations (1 stream, 0°, 30°, and 90°) for a total of 864 trials (half with a repeat and half without). The across-stream condition did not contain the 1-stream condition so the listeners completed a total of 648 trials (324 with repeat and 324 without repeat).

2. Listeners

Twenty native speakers of American English participated in this experiment. The results from four were discarded because their results in the one-stream condition were not significantly above chance. Thus, 16 listeners, 7 males and 9 females aged between 18 and 32 years (median = 21), participated. They all had normal hearing (audiometric thresholds better than 20 dB HL between 250 and 8000 Hz) and were paid for their participation. None of the listeners had previously participated in any of the previous experiments. All listeners provided written informed consent and the protocol was approved by the Institutional Review Board of the University of Minnesota.

C. Results

The d' scores, averaged across the listeners, are shown in Fig. 9. The left and right panels correspond to the within- and across-sequence tasks, respectively. For the within-sequence task, the d' scores increased in each condition (SPEC, ILD, ITD, and ALL) as the simulated separation between the sequences increased, in line with the streaming expectations. A two-way repeated-measures ANOVA with the condition (SPEC, ILD, ITD, and ALL) and angle (0°, 30°, and 90°) as within-subject factors revealed a significant effect of the condition [$F(3,45) = 16.84$, $p < 0.001$] as well as the angle [$F(2,30) = 31.9$, $p < 0.001$]. The interaction was also significant [$F(6,90) = 3.21$, $p = 0.007$]. Given the

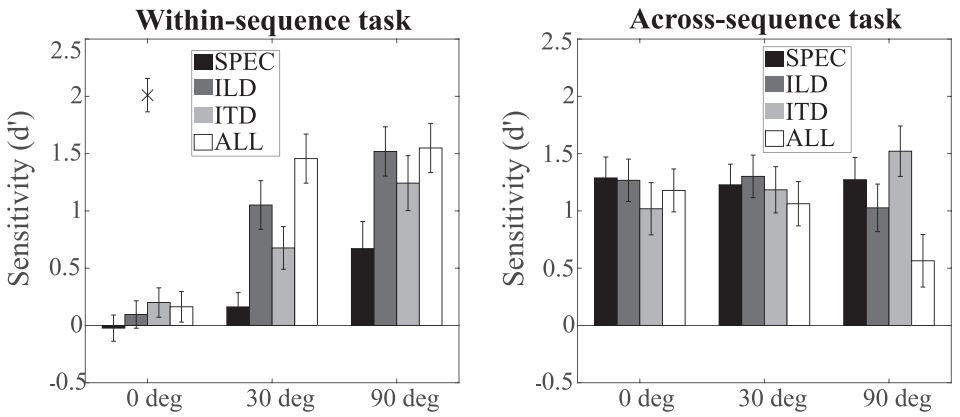

FIG. 9. Mean performance in terms of d' scores for the within- (left panel) and across- (right panel) sequence tasks for each condition tested in experiment 4. The cross in the left panel corresponds to performance in the 1-stream condition. The graded shading of the bars from black to white corresponds to the SPEC, ILD, ITD, and ALL conditions, respectively.

interaction, one-way ANOVAs were carried out for each condition separately. These ANOVAs confirmed a significant effect of simulated spatial separation for all four conditions ($p < 0.006$ in all cases), confirming that each of the spatial cues in isolation (including monaural spectral cues) provided sufficient cues for some segregation. Overall, both ITD and ILD cues seem to provide sufficient cues for robust streaming, but all three cues together provided the largest effects.

The across-sequence task is a measure of obligatory stream segregation that relies on the ability to fuse the alternating voices into one single stream. Overall, it is apparent that listeners were successfully able to ignore differences in individual spatial cues to perform the task: only the condition with ALL spatial cues present resulted in a marked decrease in performance with increasing spatial separation. A two-way repeated-measures ANOVA revealed no main effect of simulated spatial separation [$F(2,30) = 0.4$, $p = 0.67$], but a main effect of condition [$F(3,45) = 3.57$, $p = 0.021$], and a significant two-way interaction [$F(6,90) = 3.64$, $p = 0.003$], indicating that the effect of spatial separation depended on the condition tested.

Given the significant interaction, one-way repeated-measures ANOVAs were carried out on each condition separately. The ANOVAs found no significant effect of spatial separation for either SPEC [$F(2,30) = 0.048$, $p = 0.953$] or ILD [$F(2,30) = 1.339$, $p = 0.277$]. For the ITD condition, performance was found to improve significantly with increasing angle [$F(2,30) = 3.84$, $p = 0.033$]; we have no explanation for this seemingly anomalous result of improved average performance, other than to note that the effect is small and was only observed in 7 of the 16 listeners. When all the cues were present (ALL condition), there was a significant decrease in performance with increasing spatial separation [$F(2,30) = 4.02$, $p = 0.028$], in line with expectations based on streaming; this pattern of results was observed in 10 of the 16 listeners.

D. Discussion

The relative importance of binaural and spectral cues for spatially based stream segregation has been a matter of debate. Schwartz *et al.* (2012) argued that since spectral differences are encoded peripherally, they could induce more segregation than binaural cues which are processed at higher levels (Tollin, 2003; Joris and Yin, 2007). Other studies, involving sequential sounds, have shown that binaural cues tend to induce more segregation than spectral differences (Middlebrooks and Onsan, 2012), and have demonstrated the importance of ITDs over ILDs (Bremen and Middlebrooks, 2013; David *et al.*, 2015). Our results suggest that monaural spectral cues provide the weakest streaming cues for a voluntary streaming task (within-sequence task) in the horizontal plane, but that ILD and ITD cues provide comparable information, and that the combination of all three provides the strongest cues. For our obligatory streaming task (across-sequence task), only the condition with all spatial cues led to a significant effect of spatial separation that was in line with the expectations of streaming. The relatively weak effect of binaural cues in obligatory streaming tasks is consistent with findings of Oxenham (2000), using a gap-detection task, where differences in ITDs alone between the markers were not sufficient to elevate gap detection thresholds, whereas monaural level differences were.

VI. CONCLUSIONS

Several conclusions may be drawn from the present study. First, listeners are able to discriminate differences in the simulated positions of CV speech tokens in both horizontal and median planes, despite the large spectro-temporal variability between the tokens. Differences of 5° and 10° are discriminable in the horizontal and median planes, respectively. Second, listeners can extract the spectral regularities induced by a difference in position to segregate speech sounds. Third, the performance of listeners in the streaming task was based on the whole CV token and not only on the consonant or the vowel, despite the fact that the spectral cues associated with location changes in the median plane were restricted to high frequencies. Finally, in the horizontal plane, adding one binaural cue (ILD or ITD) induced more segregation than when only spectral cues were available. Despite the spectro-temporal variability in the stimuli, when all spatial cues are present, they can be sufficiently strong to produce obligatory stream segregation that prevents listeners from integrating sequences of alternating syllables into a single stream.

ACKNOWLEDGMENTS

This work was supported in part by National Institutes of Health Grant No. R01 DC007657 (A.J.O.) and by LabEX

CeLyA ANR-10-LABX-0060/ANR-11-IDEX-0007 (M.L., N.G.).

[1]Because of a programming error, only the hit rate, and not the false-alarm rate, was recorded in the single-sequence conditions of experiments 2 and 3. However, we observed that the bias, calculated as $c = -(1/2)\,(z(H) + z(FA))$, was quite constant across all other conditions (c = –0.49, SD = 0.08 and c = –0.048, SD = 0.04, for the horizontal and median plane, respectively). Based on this observation, we estimated d' in the single-sequence conditions of experiments 2 and 3 by assuming that the bias was the same as the mean of the bias in the remaining conditions. The d' was therefore calculated as $d' = 2.z(H) + 2.c$.

[2]As for experiment 2, the d' for the single-sequence condition was estimated based on the mean bias (c = –0.46, SD = 0.08 and c = –0.41, SD = 0.10, for the horizontal and median plane, respectively).

Anstis, S., and Saida, S. (**1985**). "Adaptation to auditory streaming of frequency-modulated tones," J. Exp. Psychol. **11**(3), 257–271.

Blauert, J. (**1997**). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA).

Boersma, P., and Weenink, D. (**2017**). "Praat: Doing phonetics by computer (version 6.0.31) [computer program]," http://www.praat.org/ (Last viewed 21 August 2017).

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

Bremen, P., and Middlebrooks, J. C. (**2013**). "Weighting of spatial and spectro-temporal cues for auditory scene analysis by human listeners," PloS One **8**(3), e59815.

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

Culling, J. F., and Mansell, E. R. (**2013**). "Speech intelligibility among modulated and spatially distributed noise sources," J. Acoust. Soc. Am. **133**(4), 2254–2261.

Darwin, C. J., and Hukin, R. W. (**1999**). "Auditory objects of attention: The role of interaural time differences," J. Exp. Psychol. **25**(3), 617–629.

David, M., Lavandier, M., and Grimault, N. (**2014**). "Room and head coloration can induce obligatory stream segregation," J. Acoust. Soc. Am. **136**(1), 5–8.

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

David, M., Lavandier, M., Grimault, N., and Oxenham, A. J. (**2017**). "Sequential stream segregation of voiced and unvoiced speech sounds based on fundamental frequency," Hear. Res. **344**, 235–243.

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.

Füllgrabe, C., and Moore, B. C. J. (**2012**). "Objective and subjective measures of pure-tone stream segregation based on interaural time differences," Hear. Res. **291**(1–2), 24–33.

Gardner, W. G., and Martin, K. D. (**1995**). "HRTF Measurements of a KEMAR," J. Acoust. Soc. Am. **97**(6), 3907.

Glasberg, B. R., and Moore, B. C. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**(1–2), 103–138.

Gockel, H., Carlyon, R. P., and Micheyl, C. (**1999**). "Context dependence of fundamental-frequency discrimination: Lateralized temporal fringes," J. Acoust. Soc. Am. **106**(6), 3553–3563.

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.

Joris, P., and Yin, T. C. (**2007**). "A matter of time: Internal delays in binaural processing," Trends Neurosci. **30**(2), 70–79.

Kidd, G., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (**2005**). "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**(6), 3804–3815.

Macmillan, N. A., and Creelman, D. C. (**2004**). *Detection Theory: A User's Guide*, 2nd ed., edited by Neil A. Macmillan and C. Douglas Creelman (Cambridge University Press, New York).

Martin, R. L., Mcanally, K. I., Bolia, R. S., Eberle, G., and Brungart, D. S. (**2012**). "Spatial release from speech-on-speech masking in the median sagittal plane," J. Acoust. Soc. Am. **131**(1), 378–385.

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

Middlebrooks, J. C., and Green, D. M. (**1991**). "Sound localization by human listeners," Annu. Rev. Psychol. **42**, 135–159.

Middlebrooks, J. C., and Onsan, Z. A. (**2012**). "Stream segregation with high spatial acuity," J. Acoust. Soc. Am. **132**(6), 3896–3911.

Oxenham, A. J. (**2000**). "Influence of spatial and temporal coding on auditory gap detection," J. Acoust. Soc. Am. **107**(4), 2215–2223.

Perrott, D. R., and Pacheco, S. (**1989**). "Minimum audible angle thresholds for broadband noise as a function of the delay between the onset of the lead and lag signals," J. Acoust. Soc. Am. **85**(6), 2669–2672.

Perrott, D. R., and Saberi, K. (**1990**). "Minimum audible angle thresholds for sources varying in both elevation and azimuth," J. Acoust. Soc. Am. **87**(4), 1728–1731.

Rakerd, B., Hartmann, W. M., and McCaskey, T. L. (**1999**). "Identification and localization of sound sources in the median sagittal plane," J. Acoust. Soc. Am. **106**(5), 2812–2820.

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**(5 Pt. 1), 2074–2085.

Sach, A. J., and Bailey, P. J. (**2004**). "Some characteristics of auditory spatial attention revealed using rhythmic masking release," Percept. Psychophys. **66**(8), 1379–1387.

Schwartz, A., McDermott, J. H., and Shinn-Cunningham, B. G. (**2012**). "Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences," J. Acoust. Soc. Am. **132**(2), 357–368.

Stainsby, T. H., Fu, C., Flanagan, H. J., Waldman, S. K., and Moore, B. C. J. (**2011**). "Sequential streaming due to manipulation of interaural time," J. Acoust. Soc. Am. **130**(2), 904–914.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**(3), 455–462.

Tollin, D. J. (**2003**). "The lateral superior olive: A functional role in sound source localization," Neuroscientist **9**(2), 127–143.

van Noorden, L. (**1975**). "Temporal coherence in the perception of tone sequences," Ph.D. thesis, Institute for Perception Research, University of Technology Eindhoven, Eindhoven, the Netherlands.

Wightman, F. L., and Kistler, D. J. (**1992**). "The dominant role of low-frequency interaural time differences in sound localization," J. Acoust. Soc. Am. **91**(3), 1648–1661.