Segregation of voices with single or double fundamental frequencies

Mickael L. D. Deroche, and Vincent L. Gracco

Citation: The Journal of the Acoustical Society of America **145**, 847 (2019); doi: 10.1121/1.5090107
View online: https://doi.org/10.1121/1.5090107
View Table of Contents: https://asa.scitation.org/toc/jas/145/2
Published by the Acoustical Society of America

Segregation of voices with single or double fundamental frequencies

Mickael L. D. Deroche[1,a),b)] and Vincent L. Gracco[2,a)]
[1]*Centre for Research on Brain, Language and Music, McGill University, 3640 rue de la Montagne, Montreal, H3G 2A8, Canada*
[2]*Haskins Laboratories, 300 George Street, New Haven, Connecticut 06511, USA*

(Received 9 August 2018; revised 18 January 2019; accepted 23 January 2019; published online 15 February 2019)

In cocktail-party situations, listeners can use the fundamental frequency (*F*0) of a voice to segregate it from competitors, but other cues in speech could help, such as co-modulation of envelopes across frequency or more complex cues related to the semantic/syntactic content of the utterances. For simplicity, this (non-pitch) form of grouping is referred to as "articulatory." By creating a new type of speech with two steady *F*0s, it was examined how these two forms of segregation compete: articulatory grouping would bind the partials of a double-*F*0 source together, whereas harmonic segregation would tend to split them in two subsets. In experiment 1, maskers were two same-male sentences. Speech reception thresholds were high in this task (vicinity of 0 dB), and harmonic segregation behaved as though double-*F*0 stimuli were two independent sources. This was not the case in experiment 2, where maskers were speech-shaped complexes (buzzes). First, double-*F*0 targets were immune to the masking of a single-*F*0 buzz matching one of the two target *F*0s. Second, double-*F*0 buzzes were particularly effective at masking a single-*F*0 target matching one of the two buzz *F*0s. As a conclusion, the strength of *F*0-segregation appears to depend on whether the masker is speech or not. https://doi.org/10.1121/1.5090107

[VB] Pages: 847–857

I. INTRODUCTION

A cocktail-party situation (Cherry, 1953) refers to a scene where many people talk at once in a noisy room. In such situations, the message of a target voice recorded at the listener's ears is highly masked or degraded by surrounding noise. This is partly because, contrary to our eyes which cannot look behind, our ears do not limit the auditory field, meaning that all audible sounds ultimately enter our ears. However, when keeping up a conversation and concentrating on a target talker, it seems as though one can make the sounds in the background fade away. In other words, the human brain has developed more sophisticated ways to segregate sounds that we want to listen to from the background. Cherry (1953) identified several cues that could support this perceptual separation: spatial location (voices coming from different directions), visual cues (from the lips or gestures assisting communication), speech characteristics (gender, pitch, speech rate, and accent), and transitional probabilities (linguistic and syntactic factors). The present study is concerned with perceptual separation on the basis of voice pitch cues and how it competes with other forms of grouping in speech, referred to as *articulatory cues*, regardless of the degree of linguistic processing that these cues represent. Binaural cues were removed by presenting diotic sounds over headphones (perceived inside the head, at the center), and no visual stimuli were presented. Speech characteristics concerning gender, speech rate, and accent were controlled by using material from a single male talker: listeners could hear several sentences simultaneously (up to three) but all spoken in the same manner by the same person. In such a scenario, analysis of the auditory scene (Bregman, 1990) is reduced to *fundamental frequency* (*F*0, the acoustic correlate of voice pitch) and the *articulatory content* of the utterances (i.e., the pattern of envelope modulations across frequency) eventually giving rise to a phonemic, lexical, and syntactic meaning. All *F*0 contours were flattened to gain further control over the voice's harmonic structure, even though this choice reduced the ecological relevance of the data.

It is common to differentiate simultaneous from sequential segregation. Note that the term "segregation" may often be—as in the current study—substituted for "grouping" as it simply depends on whether one views this process as splitting or binding. In the case of *F*0 cues, simultaneous segregation would refer to mechanisms by which, at a given instant or within a short time window, two harmonic structures based on different *F*0s form distinct auditory objects and get processed independently (e.g., Scheffers, 1983; Assmann and Summerfield, 1990; de Cheveigné, 1993; Culling and Darwin, 1993, 1994; de Cheveigné *et al.*, 1995; de Cheveigné *et al.*, 1997a; de Cheveigné *et al.*, 1997b). Sequential grouping would refer to the ability to track the *F*0 contour of a naturally intonated utterance and bind the discrete elements over time (Darwin and Bethell-Fox, 1977; Darwin and Hukin, 2000; Darwin *et al.*, 2003; Drullman and Bronkhorst, 2004; Vestergaard *et al.*, 2009; Clarke *et al.*, 2014). In the case of articulatory cues, one may consider simultaneous grouping based, for example, on the onset and

[a)]Also at: School of Communication Sciences and Disorders, 2001 Avenue McGill College, Montreal, H3A 1G1, Canada.
[b)]Electronic mail: mickael.deroche@mcgill.ca

offset of frequency components delimiting the start and end of a vowel or a consonant (Scheffers, 1983, Chap. III). One may consider sequential grouping as the advantage that listeners take from the context of the utterance. This may be the expectation of a verb after a noun, the ability to restore a word embedded in noise, or predict a word (or eliminate phonologically similar words) based on the semantic content of the beginning of the sentence (Lieberman, 1963; Kidd *et al.*, 2008).

Currently, it is not well understood how sequential grouping depends upon (or interacts with) the mechanisms of simultaneous grouping. Some results suggest that they may be tightly connected. For example, the hypothetical accounts of harmonic segregation of speech work best when *F*0 is fixed over time (Deroche and Culling, 2011a; Deroche *et al.*, 2014a,b; Deroche *et al.*, 2014c; Leclère *et al.*, 2017), and this could perhaps explain why perceptual continuity of this cue (along with that of vocal tract length, VTL) is so effective at guiding selective attention (Best *et al.*, 2008; Maddox and Shinn-Cunningham, 2012). On the contrary, other results suggest that they are rather independent. For example, discontinuities in *F*0 or VTL do not impair the phonemic restoration benefit, although they clearly reduce the listeners' ability to form a stream (Clarke *et al.*, 2014). Also, cueing listeners to (or, on the contrary, misleading listeners about) the target voice prior to a trial has negligible effect on speech recognition (Helfer and Freyman, 2009; Deroche *et al.*, 2017b). Finally, theories on simultaneous grouping, e.g., harmonic cancellation (de Cheveigné, 1993), rely primarily on characteristics of the masker. In contrast, sequential grouping is clearly influenced by characteristics of the target voice (Freyman *et al.*, 2004; Kidd *et al.*, 2008). Thus, to integrate these apparent inconsistencies into a comprehensive picture, there is a need to reduce complex listening situations to simpler forms of auditory masking.

However, aiming for reductionist designs is problematic: when researchers attempt to isolate a specific segregation cue, the segregation mechanisms appear to be imperfect. For example, Deroche and Culling (2013) presented a voice against a speech-modulated buzz. The term "buzz" refers to a broadband complex tone filtered to have the same spectral profile as the target material, and "speech-modulated" means that its broadband temporal envelope was not stationary but modulated by the envelope fluctuations of speech. When flattening the *F*0 of the voice to match that of the buzz, the two sources share a degree of similarity. Yet, they are perceptually distinct and certainly more distinct than when the monotonized target voice is presented against other monotonized voices. The speech reception threshold (SRT) was –5.5 dB in the former case, and increased to 0 dB in the latter. The auditory system must be capable of segregating a buzz from a voice even with a common *F*0 because there is something about the buzz that is not identified as speech. Perhaps it is the fact that it lacks spectral variations or transitions between periodic and aperiodic sounds (Chen *et al.*, 2012). When a difference in *F*0 ($\Delta F0$) is introduced between the voice and the buzz, a 4–5-dB masking release is obtained (Deroche and Culling, 2013). This is evidence that the auditory system could not *fully* separate the buzz from the voice as long as they shared the same harmonic structure, and there is no reason to believe that the two sources would be *fully* segregated with distinct *F*0s. For example, SRT decreased even further by adding a spatial location difference (in addition to $\Delta F0$s) to segregate a voice from a buzz (Leclère *et al.*, 2017). Therefore, one may justifiably wonder whether there is a point at which two auditory objects cannot be further separated, and how segregation processes facilitate or compete with each other.

We propose to use double-*F*0 speech to investigate this problem. When a monotonized voice is duplicated at two different *F*0s, one hears a sentence spoken at two different pitches, but there is nonetheless one phoneme at a time. This situation bears some similarity to polyphonic music, where *F*0 and onset/offset cues compete with each other (Rasch, 1978). We want to determine whether the auditory system would consider this artificial type of speech as an indivisible entity grouped on the basis of articulation, or two independent harmonic structures that simply happen to provide redundant speech information. The first alternative would seem plausible given the phenomenon of co-modulation masking release (CMR; Hall *et al.*, 1984; Schooneveldt and Moore, 1989). In the classical version of this phenomenon, listeners are asked to detect a tone (the target) masked by an amplitude modulated noise band. The target is better detected when presenting an additional noise band in a remote frequency channel coherently modulated with the first. The idea is that the remote noise "captures" the noise located at the target channel because they share a common property, in this case, their amplitude modulation. The two noise bands group together perceptually, thereby releasing the tone from masking. Now, it is far from trivial to extrapolate the CMR phenomenon to broadband sources, which need to be integrated across many spectral channels, but one could imagine that a mechanism would group partials from coherently modulated envelopes across channels, disregarding whether the partials were harmonically related to each other. The second alternative—that double-*F*0 speech would be processed as two independent harmonic sources disregarding co-modulation between partials—is also plausible because listeners are impressively good at segregating simultaneous steady-state vowels based on different *F*0s (de Cheveigné *et al.*, 1995; de Cheveigné *et al.*, 1997a; de Cheveigné *et al.*, 1997b). One interpretation for this ability is that the auditory system can tune to a given *F*0 and cancel its entire harmonic structure (de Cheveigné, 1993; Deroche and Culling, 2011b). This raises an interesting question: what happens to a double-*F*0 masking voice when it gets suppressed by such harmonic cancellation? Does it simply get stripped of half of its partials or is it possible that in the process of canceling one *F*0, the other one gets canceled as well because both are grouped on the basis of their common articulation?

The present study uses a parametric design with three manipulations of a target voice (low *F*0, high *F*0, or both) and three manipulations of a harmonic masker (low *F*0, high *F*0, or both) as illustrated in Fig. 1. The low and high *F*0 were separated by 8 semitones. For simplicity, the *F*0 manipulations are referred to as "L" (low), "H" (high), or "B" (both), and the source is referred to as "T" (target) or "M" (masker). The conditions of no-$\Delta F0$ (ML-TL, MH-TH, and

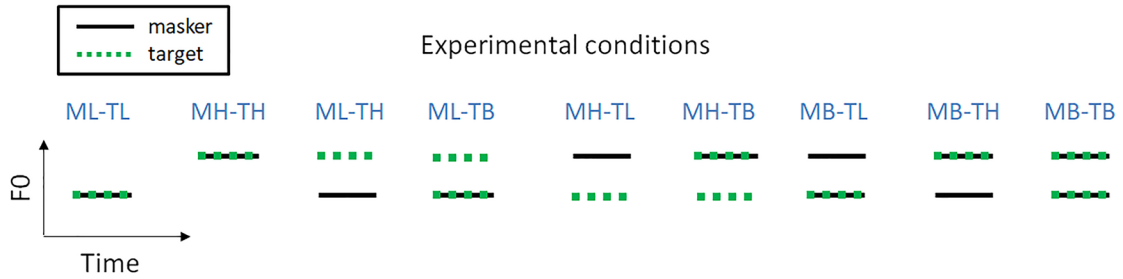

FIG. 1. (Color online) Parametric design of the nine experimental conditions, resulting from three manipulations of the target *F*0 (low, L; high, H; or both, B) and three manipulations of the masker *F*0 (L,H,B). The target is referred to as "T" and the masker as "M."

MB-TB) and the Δ*F*0 conditions between single-*F*0 sources (ML-TH and MH-TL) were used as references to examine the conditions of interest: ML-TB and MH-TB for double-*F*0 targets, as well as MB-TL and MB-TH for double-*F*0 maskers. For example, in condition ML-TB, one could expect a process akin to harmonic cancellation to tune to the single *F*0 of the masker, and the point is to determine whether the part of the target voice that coincides with it (TL) would get canceled along the way or would somehow be shielded by its common modulation with the other part (TH). Similarly, in condition MB-TL, one could expect the cancellation process to tune to the part of the masker that overlaps spectrally less with the target (MH), and the point is to determine whether ML would be left in the residual or be partly canceled with its other half because it was coherently modulated with it. Finally, this design was performed in two versions: one with speech maskers (experiment 1) and one with speech-modulated buzz maskers (experiment 2). The auditory system might use a sort of "speech-sensitive module" (Liberman and Mattingly, 1989) to set apart the voice from a buzz, regardless of whether they have one or two, common or distinct, *F*0s. However, such a module would likely struggle with speech maskers, unless advanced forms of grouping existed at a syntactical or grammatical level (Lieberman, 1963).

II. GENERAL METHODS

A. Listeners

There were 27 listeners in each experiment, and 23 of them participated in both. The number of listeners was chosen to cover three rotations of the nine experimental conditions, necessary to counterbalance the effect of speech material. All 31 listeners (9 males and 22 females) were young adults, university students for the most part, aged between 18.2 and 34.5 yr (mean of 23.4 with a standard deviation of 4.6 yr). All listeners were native speakers of North-American English and had audiometric thresholds less than 20 dB hearing level at octave frequencies between 250 Hz and 8 kHz. All subjects provided informed consent in accordance with the Institutional Review Board at McGill University, and were compensated $15 per experiment.

B. Stimuli

All target stimuli were IEEE (Institute of Electrical and Electronics Engineers) sentences (Rothauser *et al.*, 1969) spoken by a male speaker of North-American English with an original *F*0 of 88 Hz on average. There were 90 sentences in experiment 1 and 90 different sentences in experiment 2, split into blocks of 10 sentences. In addition, there were 30 sentences reserved for practice, different from those used in testing, but part of the same original material. The Praat PSOLA package (Boersma and Weenink, 2013) was used to resynthesize each sentence with a fixed *F*0 at either 100 Hz or 8 semitones above (158.7 Hz). This *F*0 manipulation induces a slight spectral tilt, such that the excitation level in unresolved regions differs somewhat as a function of *F*0. To correct for this undesirable artifact, each monotonized sentence was filtered (following the procedure described extensively in the appendix of Deroche *et al.*, 2014b) such that their excitation pattern would vary exclusively in resolved regions at equal root-mean-square (RMS) level. The target stimuli with duplicated *F*0 were simply created by adding the two *F*0-manipulated versions of the same sentence and equalizing the resulting signal to the original RMS level. Note that the excitation pattern of the three target manipulations (0 semitone, 8 semitones, 0 and 8 semitones, relative to 100 Hz) is similar to that of the two-same-male maskers (shown in Fig. 2), except that the spectral envelope of the targets differed from that of the maskers as it came from a different set of sentences.

Two types of masker were generated: speech maskers in experiment 1 and non-speech maskers in experiment 2. Speech maskers came from 18 sentences, different from any of the targets but part of the same original material. They were processed in the same way as targets, and then added in pairs and re-equalized to create two-same-male maskers. The reason to use two-sentence rather than one-sentence maskers is that listeners are known to be good at "listening in the dips" of a single interfering sentence (Hawley *et al.*, 2004; Cooke, 2006; Collin and Lavandier, 2013). If the speech reception task is done at those specific times where there is little masking energy, then *F*0-segregation would be of little use. With two sentences, the *F*0 pattern of the masker is more continuous. Moreover, two sentences are thought to result in more informational masking than one sentence (Hawley *et al.*, 2004). If articulatory grouping played a considerable role in cases where there is ambiguity about the voice to attend to, it would be better revealed with two masking sentences than with one. Thus, to clarify, there were nine maskers and each one consisted of two sentences, which could be (1) both spoken at 100 Hz, (2) both spoken at

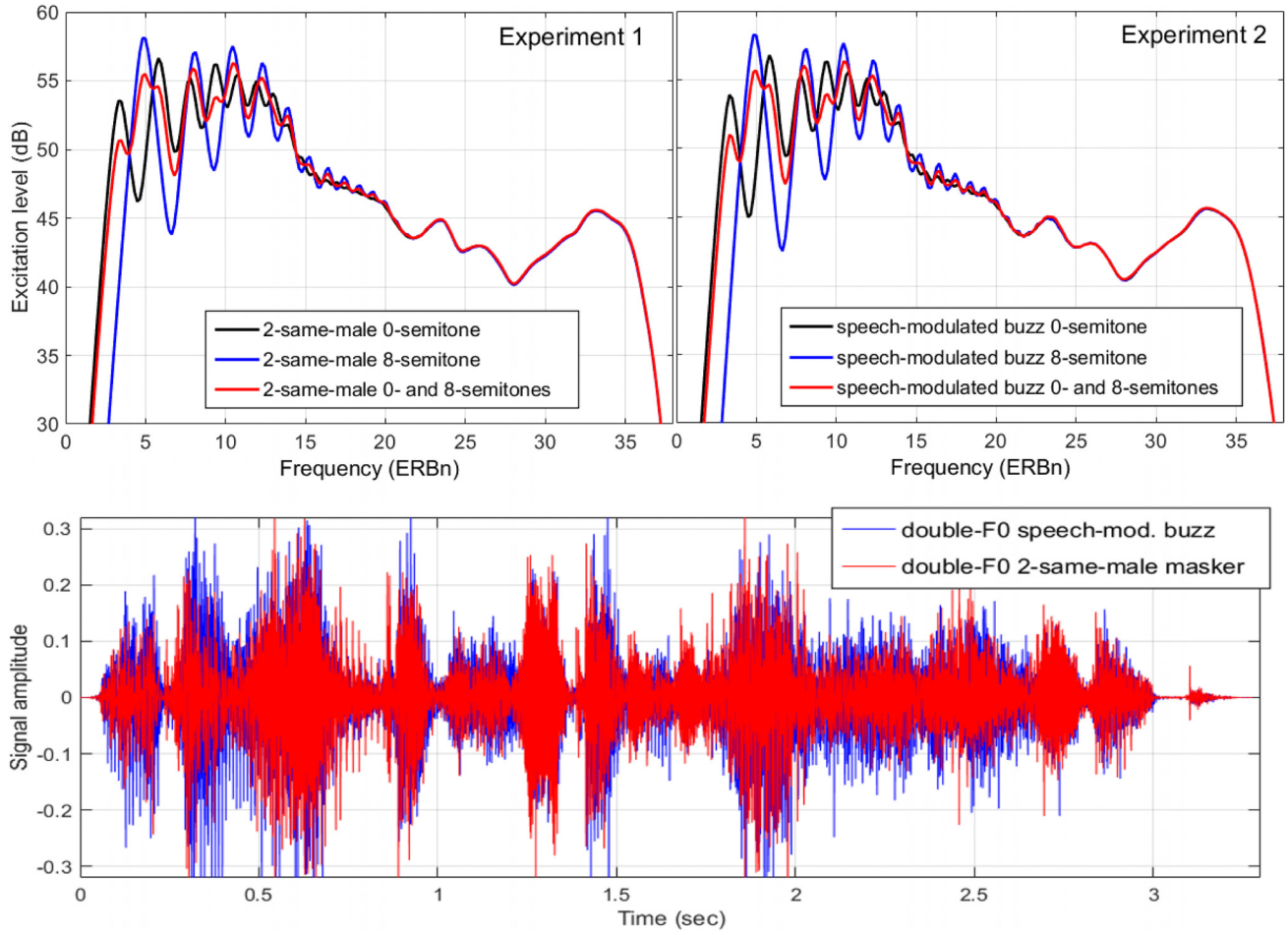

FIG. 2. (Color online) Averaged excitation patterns of the two-same-male maskers (top left) and modulated buzzes (top right) used in experiments 1 and 2, respectively, across three experimental manipulations: *F*0 at 100 Hz, *F*0 at 8 semitones above 100 Hz, and both *F*0s together. (Bottom) Broadband waveforms of the maskers (here with double *F*0, but waveforms are similar with single *F*0s), highlighting the similarity between the amplitude modulations.

158.7 Hz, and (3) both spoken at both 100 and 158.7 Hz. It was never the case that one masking sentence was spoken at 100 Hz while the other in the pair was spoken at 158.7 Hz.

Non-speech maskers were speech-modulated buzzes, modeled after the two-same-male maskers. Nine complex tones were generated from a broadband harmonic complex based on a *F*0 of 100 or 158.7 Hz and with partials in random phase. Each complex tone was then (1) filtered with a linear-phase finite impulse response filter designed to match the long-term excitation pattern of one of the two-same-male maskers, and (2) multiplied by the broadband temporal envelope of that two-same-male masker (extracted by half-wave rectification and low-pass filtering with a first-order Butterworth having a 3-dB cutoff at 40 Hz). This resulted in nine speech-modulated buzzes at 0 or at 8 semitones relative to 100 Hz. The speech-modulated buzzes with duplicated *F*0 were created by adding the two *F*0-versions of the same modulated buzz and equalizing the resulting signal to the original RMS level.

The top panels of Fig. 2 depict the excitation pattern of the three masker manipulations used in each experiment, highlighting three key points. First, the two panels are almost identical, confirming that the modulated buzzes were successfully modeled after the two-same-male maskers. Second, the excitation level across the three manipulations progressively converges into a single curve as center frequency increases. Third, the excitation pattern in resolved regions shows spectral dips and peaks, characteristics of a steady harmonic structure, with larger variations for the *F*0 at 158.7 Hz than the *F*0 at 100 Hz, and the smallest variations for the duplicated *F*0 case. On the bottom panels, the waveforms of the two masker types simply illustrate that they have similar broadband envelope fluctuations.

Note that, except for the case of duplicated *F*0, the same masker types were used in recent studies, which showed a large elevation of SRT in the case of two-same-male maskers (Deroche *et al.*, 2017b) and a strong contrast in the type of errors made, i.e., random errors for speech-modulated buzzes versus confusions with masking words for two-same-male maskers (Deroche *et al.*, 2017a). Therefore, speech maskers were expected to involve much more informational masking (Durlach *et al.*, 2003; Kidd *et al.*, 2005) than speech-modulated buzzes.

C. Procedure

The study began with three practice blocks (ten sentences each) using unprocessed targets masked by, successively, a white noise, a speech-modulated buzz, and a two-same-male masker in order to familiarize participants with the

SRT task, the interface, and the kind of maskers used in the study. The nine blocks that followed measured one SRT for each experimental condition. The order of the 90 target sentences was fixed for everyone, but the order of the conditions was rotated for successive listeners. This was necessary to counterbalance effects of order and material. SRT was measured using a one-up/one-down adaptive threshold method (Plomp and Mimpen, 1979; Culling and Colburn, 2000). The target-to-masker ratio (TMR) started at –32 dB, a level at which it is hard to even be aware of the target talker. By pressing the spacebar, listeners could repeat the first sentence, each time with a 4-dB increase in TMR. Changes in TMR occurred by adjusting the target level while keeping the maskers at 65 dB. The participants were instructed to do so until they could understand half of the first sentence. When they did, they typed their transcript (without requiring grammatical or semantical sense), which was then displayed on the screen along with the correct answer. The participants self-marked how many key words (highlighted in capitals) they obtained. The level of the target voice decreased by 2 dB if the listener had identified three, four, or five keywords correctly, and increased by 2 dB if the listener had identified two, one, or zero keywords correctly. For the nine remaining sentences in the block, it was not possible to repeat the trial. Participants heard it once, typed their transcript (or "none" if they had nothing), and self-marked their score. Measurement of each SRT was taken as the mean TMR over the last eight trials, disregarding the scores. This measurement aimed at 50% intelligibility.

Note that the task of experiment 1 was particularly difficult as participants heard three sentences simultaneously (two from the masker and one target), which were all spoken by the same male person. Without indication of which voice to track, an open-set task such as this one is almost impossible. To remedy this problem, the transcripts of each two-same-male masker were displayed on the screen, and remained constant throughout the ten sentences of a block. Listeners were instructed to focus their attention on another sentence (obviously not displayed). Although it may seem like a convoluted instruction, listeners generally tend to ignore the transcripts displayed once they have a cue to guide their attention (e.g., pitch) or once they become familiar with the semantic content of the maskers. This can be appreciated from the similarity between SRTs obtained against masking sentences (with transcripts) or a time-reversed version of them (without transcripts; Hawley *et al.*, 2004; particularly with one interferer as the comparison becomes less adequate with 2–3 interferers due to different amounts of informational masking). This can also be appreciated from the fact that listeners report words from the masking utterances despite being asked specifically not to do so (Deroche *et al.*, 2017a). Thus, displaying the masker transcripts (experiment 1) or not (experiment 2) is likely to have negligible impact relative to what the maskers were made of. Including practice, each experiment took about 90 min.

D. Equipment

Experiments were performed between July 2015 and October 2016 at the School of Communication Sciences and Disorders at McGill University inside a booth used for testing in Audiology. Signals were sampled at 44.1 kHz with a 16-bit resolution, digitally mixed, digital-to-analog converted by a sound card (Scarlett 2i4, Focusrite, Los Angeles, CA), and presented diotically over Sennheiser HD 280 headphones (Dollard-Des-Ormeaux, Quebec, Canada). The user interface was displayed on the screen of a laptop, and transcripts were typed from the laptop's keyboard.

III. EXPERIMENT 1: SPEECH MASKERS

The results of experiment 1 are illustrated in Fig. 3. The data were analyzed in four subsets using repeated-measures analyses of variance (ANOVAs) to test for several hypotheses.

A. Intrinsic intelligibility of the *F*0-manipulated voice

First, a one-way ANOVA analyzed conditions ML-TL, MH-TH, and MB-TB to test for differences in the intrinsic intelligibility of the target voice. Indeed, the mere *F*0 manipulation of a sentence can change its intelligibility (see experiment 1 of Deroche *et al.*, 2014b). For example, here, vowel formants were better defined with the double-*F*0 voice

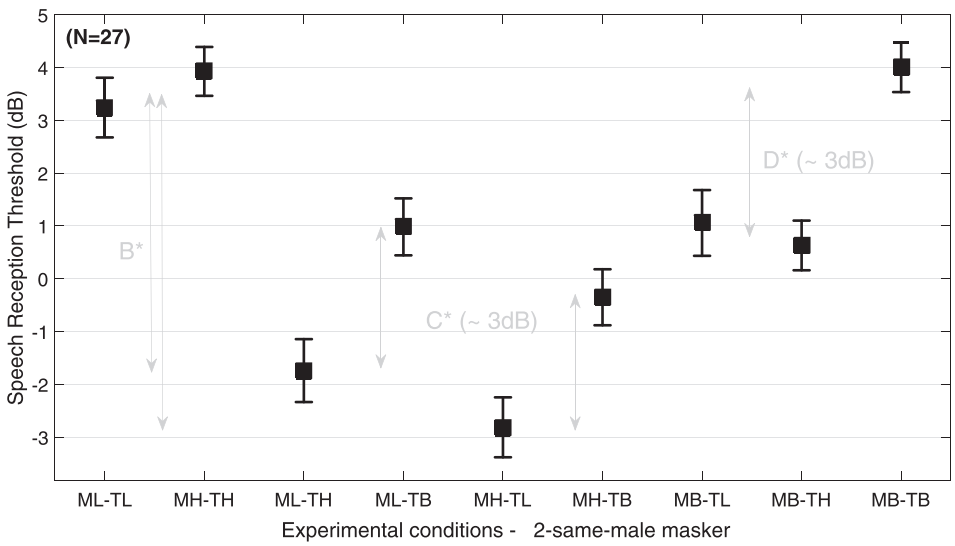

FIG. 3. Mean SRTs measured in experiment 1. The arrows and asterisks correspond to the significant effects or interactions revealed by each statistical analysis (with letters corresponding to the respective subsections).

(given that it has more spectral density) than the single-$F0$ voice, so one could have speculated a slight enhancement in the intrinsic intelligibility of the double-$F0$ voice. On the other hand, a single voice with two pitches is rather unheard of, so one may equally have speculated a slight reduction in the intrinsic intelligibility of the double-$F0$ voice on this basis. Therefore, this first test focused on the impact of the $F0$ manipulation on the intelligibility of the target voice, regardless of auditory masking. It was not the focus of the study but was necessary to factor in.

Among the three conditions (ML-TL, MH-TH, and MB-TB) when target and masker shared exactly the same $F0$s, there was no difference in SRT [$F(2,52) = 1.3$, $p = 0.287$] and none of the pairwise comparisons was significant ($p > 0.528$; Bonferroni-adjusted for multiple comparisons). Because $F0$-segregation is prevented in all these cases (or if it is somehow engaged, it does not help in separating the target information from the maskers), this lack of difference in SRT suggests that TL, TH, and TB were roughly equally intelligible.

B. $\Delta F0$ benefit

Second, a two-way ANOVA analyzed conditions ML-TL, MH-TH, ML-TH, and MH-TL to test for the presence of a masking release due to the $\Delta F0$ and whether it depended on the relative configurations of target and masker $F0$. As expected, SRTs were lower when target and masker differed in $F0$ than when they shared the same $F0$ [$F(1,26) = 135.0$, $p < 0.001$], and the masking release was larger for a $\Delta F0$ of -8 semitones (MH-TL versus ML-TL, 6.1 dB; MH-TL versus MH-TH, 6.7 dB) than a $\Delta F0$ of $+8$ semitones [ML-TH versus MH-TH, 5.7 dB; ML-TH versus ML-TL, 5.0 dB; $F(1,26) = 4.7$, $p = 0.039$]. This effect of the sign of $\Delta F0$ replicates earlier findings (Deroche *et al.*, 2014b) and supports a role for spectral glimpsing: listeners could glimpse energy from the low-$F0$ target within the larger spectral dips of the high-$F0$ masker, but were less able to do so in the reverse configuration.

C. Duplication of target *F0*

Third, a two-way ANOVA analyzed conditions ML-TH, MH-TL, ML-TB, and MH-TB to examine the effect of the target $F0$s duplication and whether this depended on the masker $F0$. SRTs were lower for the high-$F0$ masker (MH-TL and MH-TB) than the low-$F0$ masker [ML-TH and ML-TB; $F(1,26) = 8.0$, $p = 0.009$], confirming the previous observation (spectral glimpsing effect). More importantly, the main effect of the target $F0$s duplication (ML-TB versus ML-TH or MH-TB versus MH-TL) was significant [$F(1,26) = 28.4$, $p < 0.001$], elevating SRTs by about 3 dB, and it did not interact with masker $F0$ [$F(1,26) < 0.1$, $p = 0.757$].

D. Duplication of masker *F0*

Fourth, a two-way ANOVA analyzed conditions ML-TL, MH-TH, MB-TL, and MB-TH to examine the effect of the masker $F0$s duplication and whether this depended on the target $F0$. SRTs decreased by about 3 dB when duplicating the masker $F0$ [MB-TL versus ML-TL or MB-TH versus MH-TH; $F(1,26) = 57.1$, $p < 0.001$], but there was no effect of target $F0$ [$F(1,26) < 0.1$, $p = 0.768$] nor interaction [$F(1,26) = 2.1$, $p = 0.164$].

E. Discussion

SRTs increased by duplicating the target $F0$, and critically, they increased by roughly 3 dB relative to the case of single-$F0$ targets. In the double-$F0$ target, half of its power comes from the TL portion and half from the TH portion. The 3-dB elevation suggests that the ML-TB task was done exclusively by listening to the TH portion, and the MH-TB task was done exclusively by listening to the TL portion. Note that the present design does not allow us to disentangle accounts based on harmonic enhancement or harmonic cancellation (de Cheveigné *et al.*, 1995). It may be that the TL or TH portions were selected and processed exclusively in later stages in the auditory pathway, but one could also imagine harmonic cancellation tuning to the masker $F0$, removing all masker energy along with the target energy that coincided with it. The residual from such a cancellation process would also be target energy with half of its power. More to the point of the experiment, there was no indication that the double-$F0$ target was perceived as a single, indivisible, entity with regard to $F0$-segregation. Rather, harmonic segregation seemed to have taken priority over the articulatory grouping of target information.

SRTs decreased by duplicating the masker $F0$, and critically, they decreased by roughly 3 dB relative to the baseline conditions. In these conditions, the target voice was never free of masking. Once again, the auditory system could have selected the harmonic structure of the target or attempted to cancel the portion of the masker that was set on a different $F0$ than the target. In both cases, the output from the harmonic selection or cancellation would be ML-TL or MH-TH in which the masker energy had lost half of its power, and consequently SRTs were simply 3 dB lower than for the baseline conditions. Therefore, there was no indication that the double-$F0$ masker was perceived as a single, indivisible, entity: the part of the masker that was set on a distinct $F0$ could not "capture" the other half to release the target from masking.

Overall, the results of experiment 1 were clear: a double-$F0$ source (be it target or masking speech) behaves just as two harmonic sources that can be selected or canceled independently of one another.

IV. EXPERIMENT 2: NON-SPEECH MASKERS

The results of experiment 2 are illustrated in Fig. 4. The data were analyzed following the same approach as in experiment 1.

A. Intrinsic intelligibility of the *F0*-manipulated voice

Among the three conditions (ML-TL, MH-TH, and MB-TB) when target and masker shared exactly the same $F0$s, there were differences in SRT [$F(2,52) = 7.2$, $p = 0.002$]. Pairwise comparisons (with Bonferroni corrections) revealed

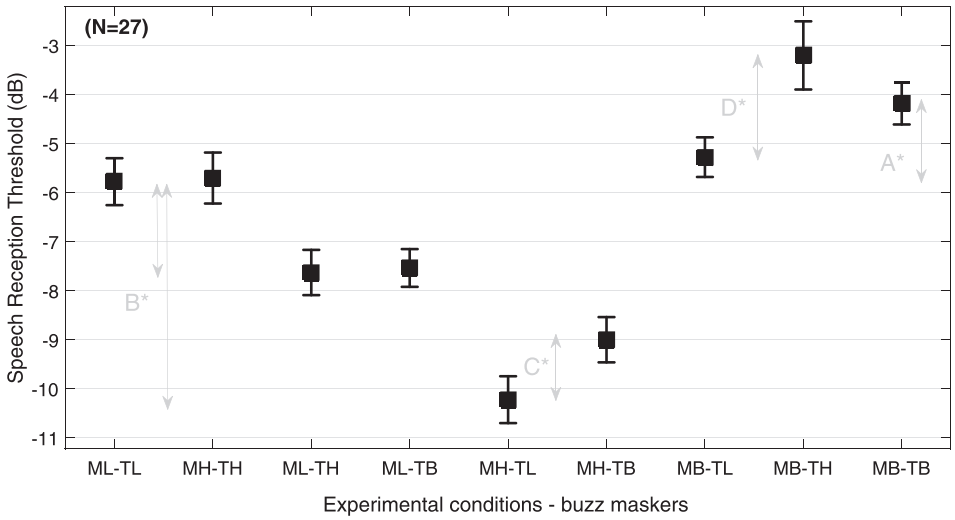

FIG. 4. Mean SRTs measured in experiment 2. The arrows and asterisks correspond to the significant effects or interactions revealed by each statistical analysis (with letters corresponding to the respective subsections).

that the MB-TB led to a higher SRT than the ML-TL and MH-TH ($p = 0.006$ and $p = 0.005$, respectively), which did not differ from each other ($p > 0.999$).

B. Δ*F*0 benefit

As expected, the second analysis (i.e., testing the presence of a Δ*F*0 benefit) revealed that SRTs were lower when target and masker differed in *F*0 than when they shared the same *F*0 [$F(1,26) = 74.1$, $p < 0.001$], and the masking release was larger for a Δ*F*0 of −8 semitones (MH-TL versus ML-TL, 4.4 dB; MH-TL versus MH-TH, 4.5 dB) than a Δ*F*0 of +8 semitones (ML-TH versus MH-TH, 1.9 dB; ML-TH versus ML-TL, 1.9 dB) [$F(1,26) = 16.6$, $p < 0.001$]. As in experiment 1, this pattern is in line with a spectral glimpsing contribution.

C. Duplication of target *F*0

In the third analysis, a significant interaction [$F(1,26) = 5.0$, $p = 0.035$] revealed that duplicating the target *F*0 elevated SRTs for the high-*F*0 masker (MH-TB versus MH-TL, $p = 0.004$), but not for the low-*F*0 masker (ML-TB versus ML-TH, $p = 0.850$). There was also a main effect of masker *F*0: SRTs were lower for the high-*F*0 masker than for the low-*F*0 masker [$F(1,26) = 32.3$, $p < 0.001$].

D. Duplication of masker *F*0

In the fourth analysis, a significant interaction [$F(1,26) = 5.3$, $p = 0.029$] revealed that duplicating the masker *F*0 lowered SRTs for the high-*F*0 target (MB-TH versus MH-TH, $p = 0.002$), not for the low-*F*0 target (MB-TL versus ML-TL, $p = 0.303$). There was also a main effect of target *F*0: SRTs were lower for the low-*F*0 target than the high-*F*0 target [$F(1,26) = 7.7$, $p = 0.010$].

E. Discussion

Perhaps, the most obvious observation is that SRTs were much lower here than in the first experiment. Specifically, SRTs in the baseline conditions (ML-TL or MH-TH) were 9–10 dB lower than in experiment 1. Although this was expected (Deroche *et al.*, 2017a; Deroche *et al.*, 2017b), this is exactly the evidence that the auditory system is capable of segregating a voice from a complex tone that yet bears some similarity to speech and without requiring any Δ*F*0. Therefore, this articulatory grouping is evidently strong and appears to override *F*0-segregation to some extent. Because buzzes generated little ambiguity about the source to attend to, there was little informational masking, which then restricted the size of the Δ*F*0 benefit (there could necessarily be little release from informational masking).

More interestingly, a number of results from experiment 2 were surprising. First, contrary to what was found earlier, the double-*F*0 voice was less intelligible than the single-*F*0 voices. It is possible that the double-*F*0 voice generated some distraction, as one never hears a sentence spoken with two simultaneous pitches on a monotone in realistic environments. So, a logical response might be to suppose that there are two talkers, and this confusion could have elevated SRTs to a small degree. But if this is so, why did we not observe this distraction effect in experiment 1? One possibility is that in experiment 1, listeners knew that they were supposed to hear several sentences, i.e., multiple pitch percepts. In fact, listening to this auditory scene (examples of stimuli are provided in the supplementary material[1]), it is quite difficult to realize that each utterance in the mixture is actually spoken at two *F*0s. The mixture simply sounds like multiple sentences with several pitches in competition. So, this listening situation is not strikingly artificial, and this is perhaps why listeners did not get confused. An alternative explanation is that there was some form of ceiling in experiment 1 where SRTs could hardly have been elevated beyond +4 dB. This ceiling issue has been acknowledged earlier in a number of studies in order to explain cases where the masking release provided by a spatial separation was unexpectedly reduced at positive TMR (Arbogast *et al.*, 2005; Freyman *et al.*, 2008; Best *et al.*, 2012) and we recently showed it to be applicable to Δ*F*0 benefits as well (Deroche *et al.*, 2017a). The idea is that a detrimental factor causes SRTs to increase (e.g., here the poorer intelligibility of TB voice compared to TL or TH, regardless of masking), but the effect is compensated by listeners taking advantage of the loudness of the

target voice at positive TMRs relative to the masking voices to effect a release from informational masking, and this is why this apparent ceiling tends to be observed with speech maskers rather than noise maskers.

More importantly, there was overall little (non-significant) effect of the target duplication. In the MH-TB, one could have expected listeners to access the TL portion with only half of its power (either from harmonic enhancement or from the residual of cancellation tuned to MH). Similarly, in the ML-TB case, one could have expected listeners to access the TH portion with half of its power. Yet, instead of a 3-dB elevation, SRT was elevated by only 1 dB in the MH-TB and not elevated at all in the ML-TB. Somehow, masking from the buzz *F*0 could be alleviated without removing the portion of the target that coincided with it. Note that this "masking immunity" of the TB voice is even more puzzling if we consider that this type of voice was intrinsically harder to understand than TL or TH (Sec. IV A). In other words, this suggests that, contrary to speech-on-speech segregation, the speech-modulated buzz and the target voice could be segregated *before* *F*0-segregation took place.

Also surprising is the fact that SRTs did not decrease by duplicating the masker *F*0. In those cases, the target voice was never free of masking, but masker energy coinciding with the target harmonic structure was reduced to half of its power. Like in experiment 1, one could have expected SRTs to be 3 dB lower than in the baseline conditions. Instead, SRT either remained at the same level or even *increased*. Somehow, the double-*F*0 buzz caused a "super-masking effect" where SRT was even *worse* than if all the masker harmonics coincided with those of the target (i.e., supposedly an ideal masker).

V. GENERAL DISCUSSION

In this study, we found that segregation of double-*F*0 targets from a single-*F*0 masker, and segregation of single-*F*0 targets from double-*F*0 maskers, does not obey the same rules when maskers are competing sentences as when maskers are speech-modulated buzzes. In experiment 1, the picture appears relatively simple. Despite the potential cues (other than *F*0) in a speech signal that binds its components together, those double-*F*0 sources can be simply halved by harmonic segregation just as though they were two independent auditory objects. In other words, we found no evidence that grouping by phonemic, semantic, syntactic, or grammatical structure (Lieberman, 1963; Freyman *et al*., 2004; Kidd *et al*., 2008; Clarke *et al*., 2014) could compete with harmonicity in speech-on-speech situations. On the other hand, against speech-modulated buzzes, a form of segregation occurs on the basis of a "speech-like identity" and it interacts with *F*0 cues in unexpected ways. First, the double-*F*0 targets appear to capture energy coinciding with the single-*F*0 buzz to prevent it from being masked. Second, the double-*F*0 maskers bolster energy coinciding with the single-*F*0 target to strengthen the masking phenomenon. This latter observation is nevertheless hard to integrate with the former: if the two harmonic structures of the double-*F*0 buzz were bound to each other, and consequently ought to be perceptually separated from the single-*F*0 target, then why did masking persist or even increase? One could have expected the reverse effect: a reduction in masking. We have no convincing account.

Any study on masked speech perception deserves a word on modulation masking. The slow envelope modulations of a target voice (below 10 Hz) are essential to its intelligibility (Drullman *et al*., 1994; Elliott and Theunissen, 2009). Maskers whose envelopes fluctuate at these low rates can therefore prevent listeners from reconstructing the articulation of the target (Jørgensen and Dau, 2011; Jørgensen *et al*., 2013). Contrary to noise that has slow and random envelope fluctuations (Stone and Moore, 2014), stationary harmonic complexes are generally considered *weak* modulation maskers because their envelopes fluctuate at the rate of *F*0 and consistently across frequency. This is no longer the case when a speech-like broadband modulation is applied on them, such as our present buzzes, but at least this modulation is common across frequency. A mixture of two masking sentences exhibits a number of local and inconsistent fluctuations that likely generate more modulation masking than for the buzzes. This account would contribute to the large elevation of SRTs observed in experiment 1. More problematically, double-*F*0 maskers have twice the spectral density of single-*F*0 maskers. In this respect, they are in fact similar to inharmonic complexes whose frequency components (at least some of them) are close to each other and produce a beating, i.e., a slow envelope modulation. As a consequence, the double-*F*0 maskers could, in principle, be expected to be stronger modulation maskers than the single-*F*0 maskers. Yet, in experiment 1, SRT for the MB-TB situation was not different from the ML-TL or MH-TH situations, and the 3-dB effects (elevation or reduction in SRTs) are well accounted for by power differences in the spectral domain rather than the modulation domain. Also, note that this interpretation would not help in explaining the difference between MB-TL and MB-TH. Thus, it is not trivial to appreciate how the present findings would be confounded by differences in modulation masking.

The idea that *F*0-segregation does not obey the same rules with speech or simpler linguistic signals is not novel. This is best exemplified by the evolution of the masking release as a function of the size of $\Delta F0$, which is drastically different for vowels and sentences. Vowel identification improves sharply for extremely small $\Delta F0$s (Scheffers, 1983; Culling and Darwin, 1993; de Cheveigné *et al*., 1997a; de Cheveigné *et al*., 1997b): differences as small as 7 cents generated some measurable benefit (de Cheveigné, 1999). In contrast, the benefit observed with sentences is thought to require several semitones (Brokx and Nooteboom, 1982; Darwin *et al*., 2003; Drullman and Bronkhorst, 2004). However, recent studies (Deroche and Culling, 2013; Deroche *et al*., 2017b) demonstrated that when speech is presented against buzzes, the masking release is almost fully obtained with a $\Delta F0$ of only 2 semitones as it is in double-vowel identification studies. So, this distinction is not due to speech *per se* but the fact that maskers are speech or not. It is thus interesting to list the differences that exist between the maskers of experiments 1 and 2. First, the spectral

envelope of buzzes is fixed, whereas that of masking sentences is dynamic (they were only matched in long-term average), which likely contributes to a speech-like identity. Second, the buzzes were modulated from their broadband temporal envelopes, whereas masking sentences have different envelope modulations at different frequencies. There is emerging evidence that *F*0-segregation interacts with the temporal modulations in the masker: Leclère *et al.* (2017) applied a one-voice broadband modulation to monotonized or intonated buzzes and found a dip-listening benefit (relative to stationary buzzes) only in the intonated case (i.e., when *F*0-segregation is disengaged). It is likely that such interactions would take place—and perhaps differently—when temporal modulations are restricted to a narrower frequency region and inconsistent across frequency. Third, buzzes are continuously voiced (except for a few silent breaks in the two-voice broadband envelope), whereas masking sentences have unvoiced portions. In principle, *F*0-segregation should have *more impact* the less interrupted the *F*0 contour. However, the alternation between voiced and unvoiced portions in a masker is particularly effective at generating informational masking (Chen *et al.*, 2012), and a $\Delta F0$ can then be used to release from informational masking. This is why, provided that $\Delta F0$s are large enough, the masking release can be larger with masking sentences than with buzzes (Deroche and Culling, 2013; Deroche *et al.*, 2017a; Deroche *et al.*, 2017b). For all these reasons, therefore, it may not be surprising that *F*0-segregation behaves differently for buzzes and speech maskers. Nevertheless, our ability to comprehend how listeners perform in speech-on-speech situations requires us to break down this complex scene into simpler situations of masking. The fact that some results seem to apply to both masker types (e.g., spectral glimpsing effects) while other results are masker-specific (e.g., interaction with articulatory cues) is a step forward in our understanding of segregation processes.

One may question the validity of the present conclusions for sources with naturally intonated *F*0s. We have recently looked at this question carefully by measuring SRT for a target voice with a monotonized or naturally intonated *F*0 against a buzz with a monotonized or naturally intonated *F*0 in all possible orthogonal manipulations (Leclère *et al.*, 2017). Instantaneous $\Delta F0$s resulting from an intonated target against a monotonized buzz, with the same *F*0 on average, provided as much masking release as a fixed 3-semitone $\Delta F0$ between monotonized sources, suggesting that mean $\Delta F0$ is *meaningless* in the case of fluctuating *F*0s. Furthermore, instantaneous $\Delta F0$s resulting from a monotonized target against an intonated buzz, with the same *F*0 on average, provided no benefit, emphasizing that periodicity in the masker is not enough: it must be sufficiently stable over time for the different underlying mechanisms (Deroche *et al.*, 2014a,b; Deroche *et al.*, 2014c) to take place. In contrast, periodicity in the target voice has repeatedly been shown to have little importance (de Cheveigné *et al.*, 1995; Deroche and Culling, 2011a; Steinmetzger and Rosen, 2015) and may therefore fluctuate without causing any impairment (in fact, it is rather beneficial to prosody). In other words, studies that control (e.g., flatten) *F*0 variations in speech are *necessary* to further our understanding of the segregation processes, but it is likely that none of the present findings would be directly observable with intonated sources because (1) the masking release would be drastically limited by fast variations in the masker *F*0 even in simple cases such as ML-TH, and (2) the elevation or reduction in SRTs would not amount exactly to 3 dB.

It is also difficult, at the moment, to generalize the present results to cocktail-party situations. Experiment 1 presented a listening situation that is highly artificial and only available in the laboratory. In natural settings, three simultaneous sentences will be spoken by different talkers. Even if they happened to be located all behind the listener (i.e., no visual cues) and very close to each other (i.e., weak binaural cues), there would be at least a difference in vocal tract (among other differences, including accent, speech rate, etc.). For a target female voice against two masking male voices, it may be that grouping based on vocal-tract length (Brungart *et al.*, 2001; Darwin *et al.*, 2003; Culling and Porter, 2005; Vestergaard *et al.*, 2009) would override *F*0-segregation to some degree, similarly to the effects reported in experiment 2. Or perhaps, the results of experiment 1 are truly generalizable to speech-on-speech situations, and in addition to binaural, visual, speech characteristics, and higher-level contextual cues (Kidd *et al.*, 2008; Clarke *et al.*, 2014), *F*0-segregation remains one of the key mechanisms to simplify the auditory scene (Bregman, 1990). Here again, further work will be needed to advance those questions.

One approach that is directly related to the present objective is the use of sine-wave analogs of speech (Bailey *et al.*, 1977; Remez *et al.*, 1981). By reproducing the frequency and amplitude contours of the first three formants with simple sinusoids, utterances can be made intelligible. Moreover, this ability depends to some degree on instructing listeners that they are about to hear speech. Without such priming, the auditory system does not readily identify such stimuli as speech. This begs the question that is central to the current study: how does the auditory system identify a given source as linguistic? If there is a "speech module" in the brain (Liberman and Mattingly, 1989), what acoustic property does it spot to make a decision about the speech-like quality of a signal? Clearly, being a broadband signal or having a harmonic structure is neither necessary (sine-wave speech is intelligible while missing both of those characteristics) nor sufficient (consider a speech-shaped complex like our buzz) to be considered as a speech-like entity. It may be that having a dynamic formant structure is the key because it emphasizes that the source passes through a range of vocal tract resonances. However, a dynamic formant structure *alone* is insufficient for segregation. To illustrate, one could argue that sine-wave speech has nothing *but* a dynamic formant structure, and once two sine-wave utterances are mixed together, listeners are not capable of segregating them (Barker and Cooke, 1999). This is intriguing because the sine-wave stimuli are so devoid of speech-like structure that it considerably removes any energetic masking confound. For example, in the second-formant competitor (F2C) paradigm, rather than presenting the three sinusoids in the same ear, *F*1 and *F*3 are presented in one ear and *F*2 in the other

ear. In the absence of a competitor, these three formants can be integrated across ears and intelligibility retrieved. However, intelligibility is substantially reduced when introducing a F2C in the same ear as $F1$ and $F3$. To explore this phenomenon further, Roberts *et al.* (2010) focused on the characteristics of the F2C: either its amplitude contour was flattened, keeping its frequency contour intact, or vice versa. They found that intelligibility suffered in the former case but not the latter. In other words, across-formant grouping relies on formant frequency modulation, but cares little about their amplitude contour. This account would appear to be consistent with the present results. Here, it is plausible that the auditory system recognized the target voice as being a speech-like entity because it possessed a dynamic formant structure. In contrast, the buzz did not; it had a fixed spectral envelope and its speech-modulated temporal envelope would have produced little interference to the across-formant grouping in charge of binding elements of the target voice. To summarize, this interpretation suggests that across-formant grouping could be responsible for segregation of sources regarded as linguistic versus non-linguistic. However, it does not explain why the double-$F0$ buzzes produced a super-masking effect. Whatever the grouping mechanisms in place, it would seem that the mere assumption that grouping mechanisms occur in a sequential order is too simplistic. Perhaps, it is more appropriate to think of a combinatorial process in which articulatory grouping and eventually linguistic grouping is itself subdivided into a number of processes, among which harmonic relations play a role.

ACKNOWLEDGMENTS

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada awarded to V.L.G. We are grateful to all participants for their time and effort.

[1]See supplementary material at https://doi.org/10.1121/1.5090107 for examples of experimental stimuli: a target sentence "the BEETLE DRONED in the HOT JUNE SUN" (SuppPubmm1x.wav), a two-same-male masker "WAKE and RISE, and STEP into the GREEN OUTDOORS; the PUP JERKED the LEASH as he saw a FELINE SHAPE" (SuppPubmm2x.wav), and a speech-modulated buzz masker (SuppPubmm3x.wav), which were manipulated to have a steady low $F0$ ($x = a$), high $F0$ ($x = b$), or both ($x = c$).

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2005**). "The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **117**, 2169–2180.

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

Bailey, P. J., Summerfield, Q., and Dorman, M. (**1977**). "On the identification of sine-wave analogues of certain speech sounds," Haskins Laboratories Status Report on Speech Research, Report No. SR-51/52, New Haven, CT, pp. 1–25.

Barker, J., and Cooke, M. (**1999**). "Is the sine-wave speech cocktail-party worth attending to?," Speech Commun. **27**, 159–174.

Best, V., Marrone, N., Mason, C. R., and Kidd G., Jr. (**2012**). "The influence of non-spatial factors on measures of spatial release from masking," J. Acoust. Soc. Am. **13**, 3103–3110.

Best, V., Ozmeral, E. J., Kopčo, N., and Shinn-Cunningham, B. G. (**2008**). "Object continuity enhances selective auditory attention," Proc. Natl. Acad. Sci. U.S.A. **105**, 13174–13178.

Boersma, P., and Weenink, D. (**2013**). "Praat: Doing phonetics by computer (version 5.3.85) [computer program]," http://www.praat.org/ (Last viewed April 1, 2013).

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

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

Brungart, D., Simpson, B., Ericson, M., and Scott, K. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Chen, J., Li, H., Li, L., Wu, X., and Moore, B. C. J. (**2012**). "Informational masking of speech produced by speech-like sounds without linguistic content," J. Acoust. Soc. Am. **131**, 2914–2926.

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

Clarke, J., Gaudrain, E., Chatterjee, M., and Baskent, D. (**2014**). "T'ain't the way you say it, it's what you say—Perceptual continuity of voice and top-down restoration of speech," Hear. Res. **315**, 80–87.

Collin, B., and Lavandier, M. (**2013**). "Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers," J. Acoust. Soc. Am. **134**, 1146–1159.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

Culling, J. F., and Colburn, H. S. (**2000**). "Binaural sluggishness in the perception of tone sequences and speech in noise," J. Acoust. Soc. Am. **107**, 517–527.

Culling, J. F., and Darwin, C. J. (**1993**). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F_0," J. Acoust. Soc. Am. **93**, 3454–3467.

Culling, J. F., and Darwin, C. J. (**1994**). "Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating," J. Acoust. Soc. Am. **95**, 1559–1569.

Culling, J. F., and Porter, J. (**2004**). "Effects of differences in the accent and gender of competing voices on speech segregation," in *Auditory Signal Processing. Physiology, Psychoacoustics and Models*, edited by D. Pressnitzer, A. de Cheveigné, S. Mc Adams, and L. Collet (Springer Verlag, New York), pp. 307–313.

Darwin, C. J., and Bethell-Fox, C. E. (**1977**). "Pitch continuity and speech source attribution," J. Exp. Psychol. Hum. Percept. Perform. **3**, 665–672.

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**, 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**, 970–977.

de Cheveigné, A. (**1993**). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. **93**, 3271–3290.

de Cheveigné, A. (**1999**). "Waveform interactions and the segregation of concurrent vowels," J. Acoust. Soc. Am. **106**, 2959–2972.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (**1997a**). "Concurrent vowel segregation. I. Effects of relative amplitude and F_0 difference," J. Acoust. Soc. Am. **101**, 2839–2847.

de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (**1995**). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. **97**, 3736–3748.

de Cheveigné, A., McAdams, S., and Marin, C. (**1997b**). "Concurrent vowel segregation. II. Effects of phase, harmonicity and task," J. Acoust. Soc. Am. **101**, 2848–2856.

Deroche, M. L. D., and Culling, J. F. (**2011a**). "Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation," J. Acoust. Soc. Am. **130**, 2855–2865.

Deroche, M. L. D., and Culling, J. F. (**2011b**). "Narrow noise band detection in a complex masker: Masking level difference due to harmonicity," Hear. Res. **282**, 225–235.

Deroche, M. L. D., and Culling, J. F. (**2013**). "Voice segregation by difference in fundamental frequency: Effect of masker type," J. Acoust. Soc. Am. **134**, EL465–EL470.

Deroche, M. L. D., Culling, J. F., and Chatterjee, M. (**2014c**). "Phase effects in masking by harmonic complexes: Detection of bands of speech-shaped noise," J. Acoust. Soc. Am. **136**, 2726–2736.

Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. (**2014a**). "Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicity," J. Acoust. Soc. Am. **135**, 2873–2884.
Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. (**2014b**). "Roles of target and masker fundamental frequency in voice segregation," J. Acoust. Soc. Am. **136**, 1225–1236.
Deroche, M. L. D., Culling, J. F., Lavandier, M., and Gracco, V. L. (**2017a**). "Reverberation limits the release from informational masking obtained in the harmonic and binaural domains," Attent. Percept. Psychophys. **79**, 363–379.
Deroche, M. L. D., Limb, C. J., Chatterjee, M., and Gracco, V. L. (**2017b**). "Similar abilities of musicians and non-musicians to segregate voices by fundamental frequency," J. Acoust. Soc. Am. **142**, 1739–1755.
Drullman, R., and Bronkhorst, A. (**2004**). "Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers," J. Acoust. Soc. Am. **116**, 3090–3098.
Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.
Durlach, N., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G., Jr. (**2003**). "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**, 368–379.
Elliott, T. M., and Theunissen, F. E. (**2009**). "The modulation transfer function for speech intelligibility," PLoS Comput. Biol. **5**, e1000302.
Freyman, R. L., Balakrishnan, U., and Helfer K. S. (**2004**). "Effects of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.
Freyman, R. L., Balakrishnan, U., and Helfer K. S. (**2008**). "Spatial release from masking with noise-vocoded speech," J. Acoust. Soc. Am. **124**, 1627–1637.
Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectro-temporal pattern analysis," J. Acoust. Soc. Am. **76**, 50–56.
Hawley, M., Litovsky, R., and Culling, J. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.
Helfer, K. S., and Freyman, R. L. (**2009**). "Lexical and indexical cues in masking by competing speech," J. Acoust. Soc. Am. **125**, 447–456.
Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**, 1475–1487.
Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**, 436–446.
Kidd, G., Jr., Best, V., and Mason, C. R. (**2008**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**, 3793–3802.
Kidd, G., Jr., Mason, C. R., and Gallun, F. J. (**2005**). "Combining energetic and informational masking for speech identification," J. Acoust. Soc. Am. **118**, 982–992.
Leclère, T., Lavandier, M., and Deroche, M. L. D. (**2017**). "The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location," Hear. Res. **350**, 1–10.
Liberman, A. M., and Mattingly, I. G. (**1989**). "A specialization for speech perception," Science **243**, 489–494.
Lieberman, P. (**1963**). "Some effects of semantic and grammatical context on the production and perception of speech," Lang. Speech **6**, 172–187.
Maddox, R. K., and Shinn-Cunningham, B. G. (**2012**). "Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention," J. Assoc. Res. Otolaryngol. **13**, 119–129.
Plomp, R., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech-reception threshold for sentences," Audiology **18**, 43–52.
Rasch, R. A. (**1978**). "The perception of simultaneous notes such as in polyphonic music," Acta Acust. Acust. **40**, 21–33.
Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–949.
Roberts, B., Summers, R. J., and Bailey, P. J. (**2010**). "The perceptual organization of sine-wave speech under competitive conditions," J. Acoust. Soc. Am. **128**, 804–817.
Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Scheffers, M. T. M. (**1983**). "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. thesis, Rijksuniversiteit Groningen, The Netherlands.
Schooneveldt, G. P., and Moore, B. C. J. (**1989**). "Comodulation masking release (CMR) as a function of masker bandwidth, modulator bandwidth, and signal duration," J. Acoust. Soc. Am. **85**, 273–281.
Steinmetzger, K., and Rosen, S. (**2015**). "The role of periodicity in perceiving speech in quiet and in background noise," J. Acoust. Soc. Am. **138**, 3586–3599.
Stone, M. A., and Moore, B. C. J. (**2014**). "On the near non-existence of 'pure' energetic masking release for speech," J. Acoust. Soc. Am. **135**, 1967–1977.
Vestergaard, M. D., Fyson, N. R. C., and Patterson, R. D. (**2009**). "The interaction of vocal characteristics and audibility in the recognition of concurrent syllables," J. Acoust. Soc. Am. **125**, 1114–1124.