Phase effects in masking by harmonic complexes: Detection of bands of speech-shaped noise

Mickael L. D. Deroche^{a)}

Department of Otolaryngology, Johns Hopkins University School of Medicine, 818 Ross Research Building, 720 Rutland Avenue, Baltimore, Maryland 21205

John F. Culling

School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, United Kingdom

Monita Chatterjee

Auditory Prostheses and Perception Laboratory, Boys Town National Research Hospital, 555 North 30th Street, Omaha, Nebraska 68131

(Received 19 October 2013; revised 4 August 2014; accepted 5 September 2014)

When phase relationships between partials of a complex masker produce highly modulated temporal envelopes on the basilar membrane, listeners may detect speech information from temporal dips in the within-channel masker envelopes. This source of masking release (MR) is however located in regions of unresolved masker partials and it is unclear how much of the speech information in these regions is really needed for intelligibility. Also, other sources of MR such as glimpsing in between resolved masker partials may provide sufficient information from regions that disregard phase relationships. This study simplified the problem of speech recognition to a masked detection task. Target bands of speech-shaped noise were restricted to frequency regions containing either only resolved or only unresolved masker partials, as a function of masker phase relationships (sine or random), masker fundamental frequency (F_0) (50, 100, or 200 Hz), and masker spectral profile (flat-spectrum or speech-shaped). Although masker phase effects could be observed in unresolved regions at F_0 s of 50 and 100 Hz, it was only at 50-Hz F_0 that detection thresholds were ever lower in unresolved than in resolved regions, suggesting little role of envelope modulations for harmonic complexes with F_0 s in the human voice range and at moderate level.

© 2014 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4896457>]

PACS number(s): 43.66.Dc, 43.66.Nm, 43.71.Gv [VMR]

Pages: 2726–2736

I. INTRODUCTION

Speech in a background masker is more easily understood when that masker has a regular harmonic structure (de Cheveigné *et al.*, 1995). This phenomenon has practical importance when the masker is a competing talker, but the underlying mechanisms are not completely understood. Models of the process have suggested perceptual grouping of harmonics from the two voices (Assmann and Summerfield, 1990; Meddis and Hewitt, 1992), or selective cancellation of the masking voice (de Cheveigné, 1997). These mechanisms would account for the fact that each voice must have a different fundamental frequency (F_0) for masking release (MR) to occur. Such models rely on analyzing the power spectrum in some way (including spectrotemporal mechanisms that apportion energy in a given channel to sources with different periods). However, masking is not entirely dependent on the power spectrum of the masking sound.

Kohlrausch and Sander (1995) and Carlyon and Datta (1997) found substantial variations in the detection threshold of a brief tone pip at different points in the period of a harmonic masker. Kohlrausch and Sander termed this variation

the masking period pattern. This pattern depended on the phase spectrum; positive Schroeder-phase complexes produced large reductions in threshold at certain points in the masker period, but negative Schroeder-phase complexes did not. Due to the phase response of the basilar membrane (BM), only responses to the positive Schroeder-phase masker are highly modulated on the BM at the signal frequency, indicating that very brief dips in masker energy were somehow exploited. Carlyon and Datta demonstrated the level-dependency of this effect, which was larger at higher masker levels, and inferred a link with the non-linear gain of the BM. It was further proposed that the non-linear gain of the BM was the most important factor since it could provide a MR in modulated maskers without assuming any selective listening in the masker dips (Alcántara *et al.*, 2003). For instance, compression of the temporal peaks in the masker might reduce forward masking of target energy in the following dips (Gockel *et al.*, 2002).

These phenomena raise the possibility that a speech signal might also be detected more easily when the phase spectrum of an interfering voice is such that within-channel envelopes are sufficiently modulated on the BM. Some evidence to support this possibility comes from Summers and Leek (1998). Using a voice masked by flat-spectrum harmonic complexes based on a F_0 of 100 Hz, they found a difference in speech reception threshold (SRT) between maskers

^{a)}Author to whom correspondence should be addressed. Electronic mail: mderoch2@jhmi.edu

in positive and negative Schroeder phase, which was also level-dependent for normal-hearing listeners. Curiously, however, phase effects in tone detection *decreased* as target level increased from 60 to 80 dB sound pressure level (SPL), whereas phase effects in speech recognition *increased* as target level increased. Given the shape of the input–output function of the BM (Yates, 1990), Summers and Leek suggested that an increase in level had shifted the pure tones up from the non-linear, compressive, region to a higher-level linear region, resulting in less differential amplification, i.e., less phase effect. In contrast, an increase in the level of sentences had shifted speech energy within any given band from the low-level linear region to the intermediate non-linear region, resulting in more differential amplification, i.e., more phase effect. Green and Rosen (2013) recently used Schroeder harmonic complexes (as well as other phase relationships) with flat-spectrum or speech-shaped spectral profiles, and a fixed or a naturally varying F_0 . The flat-spectrum maskers with fixed F_0 emulated those used by Summers and Leek, while speech-shaped maskers with intonated F_0 patterns are more relevant to interfering speech. They confirmed that masker phase effects in speech recognition were highly level-dependent, with minor or negligible effects at 60 dB SPL.

Aside from the dependency on masker level that has been largely documented, the issue of the frequency region concerned with masker phase effects has received less attention and is nonetheless key in the context of speech intelligibility. Conventionally, speech intelligibility in noise is predicted based on the signal-to-noise ratio (and therefore audibility) in each frequency channel. Channels between about 0.4 and 4 kHz contain the most useful information (ANSI, 1997). However, decent levels of speech intelligibility may be possible with signals restricted to a narrow bandwidth within this range. When considering that a SRT measurement only targets 50% intelligibility, it is unclear how much of the speech information is needed and from which frequency region. This is a critical issue to resolve to understand the extent to which masker phase effects occur in speech recognition. Phase effects must originate from regions of largely unresolved masker partials because several partials need to pass through the same critical band in order to produce highly modulated envelopes. This means that the speech cues being released from masking by this phenomenon are located in higher and higher frequencies as masker F_0 increases. Logically, there must be a F_0 at which strong modulations in the BM response produced by the masker are useless simply because the target voice would be intelligible from regions of resolved partials only. The speech cues being released from masking should also depend somewhat on the masker spectral profile, because there is more masking in high-frequency regions in the presence of flat-spectrum maskers than in the presence of speech-shaped maskers, more representative of environmental masking. Therefore, in principle, masker phase effects in speech recognition should strongly depend on masker F_0 and masker spectral profile.

Other sources of MR may also affect the extent to which phase effects transfer to speech recognition. In regions of resolved partials, listeners may be able to extract speech

cues in the channels located at the masker spectral dips, a process we refer to as *spectral glimpsing*. Contrary to the phase effects, which should be less and less beneficial as masker F_0 increases, spectral glimpsing effects should increase with F_0 , since spectral dips broaden and deepen as F_0 increases. In other words, the benefit of masker envelope modulations in speech recognition can hardly be evaluated without carefully considering how much MR occurs for speech cues located in other regions of the masker, and how useful these cues are for intelligibility.

To approach this problem, the present study compared the detectability of speech-shaped sources in resolved versus unresolved regions, for a given masker F_0 , a given spectral profile, and at a given target-to-masker (TMR) ratio. A fundamental assumption was that the availability of speech cues in a given region, and its respective contribution to intelligibility, will be reflected by the audibility of a band of noise with the same spectral shape and bandwidth as average speech in that region. Since different bands of speech-shaped noise have different energy, they have different audibility in a given masker, so it was necessary to measure their intrinsic audibility by using noise (white or speech-shaped) as a third masker type. For the TMR to be comparable across regions, the TMR always referred to the level of the full target speech-shaped noise prior to low-pass filtering in resolved regions, or high-pass filtering in unresolved regions, relative to the masker fixed at 65 dB SPL. With three masker types (sine-phase, random-phase, and noise masker), three F_0 s (50, 100, and 200 Hz), two spectral regions (resolved or unresolved), and two spectral profiles (flat-spectrum or speech-shaped), there were 36 conditions overall. First, a computational analysis modeled the use of spectral dips and envelope modulations to make some predictions regarding the MR they could provide as a function of F_0 and spectral profile. Subsequently, the experiment evaluated whether these predictions were observed and compared detection thresholds between resolved and unresolved regions to conclude on the role of strong modulations in the BM response to harmonic maskers. If detection thresholds were lower in unresolved than in resolved regions, this role should be primary especially when the region covers a substantial range of the speech spectrum. In contrast, if detection thresholds were higher in unresolved than in resolved regions, it is more likely that speech could be recovered from relatively low frequencies.

II. COMPUTATIONAL ANALYSIS

A. Stimuli and conditions

The same stimuli and conditions were used in both the computational analysis and the experiment. All target stimuli were constructed from 200-ms Gaussian noise samples, with 30-ms onset and offset ramps, passed through a linear-phase finite-impulse-response filter designed to match the excitation pattern of average speech (frequency response shown in Fig. 1). The average speech was based on a corpus of 450 sentences spoken by a male talker and the same 450 sentences spoken by a female talker, taken from the IEEE list (Rothausen *et al.*, 1969). The resulting speech-shaped noise

FIG. 1. Long-term excitation pattern of speech (averaged over a male and female speaker) compared with that of flat-spectrum broadband noise. The thin lines represent the frequency response of the low- and high-pass filters that delimited the resolved and unresolved regions of harmonic complexes based on F_0 s of 50, 100, and 200 Hz. Here, the patterns are compensated for the upwards spread of masking, for ease of comparison with the filter slopes.

was then passed through a further low- or high-pass filter depending on the experimental condition. The filter cutoffs were designed to limit the target bandwidth to spectral regions that contained only resolved or unresolved masker partials.

According to the definition proposed by Shackleton and Carlyon (1994), partials are considered resolved when the filter passes fewer than 2 partials and considered unresolved when the filter passes more than 3.25 partials within its 10-dB-down bandwidth. With rounded-exponential auditory filters (Glasberg and Moore, 1990), partials of a 50-Hz F_0 complex are, by this definition, resolved below a center frequency of 286 Hz and unresolved beyond a center frequency of 609 Hz. Partial of a 100-Hz F_0 complex are resolved below a center frequency of 802 Hz and unresolved beyond a center frequency of 1447 Hz. Partial of a 200-Hz F_0 complex are resolved below a center frequency of 1833 Hz and unresolved beyond a center frequency of 3122 Hz. Butterworth sixth-order filters with slopes of -30 dB per octave were used to low-pass filter the speech-shaped noise in the resolved regions or high-pass filter the speech-shaped noise in the unresolved regions. Figure 1 illustrates the frequency response of these six filters (thin lines) compared to gender-averaged speech energy (thick lines). There were therefore six target bands (2 spectral regions $\times$ 3 F_0 s).

The maskers had three forms of excitation: sine phase (SP), random phase (RP) harmonic complexes, and noise. The complexes consisted of 440, 220, or 110 harmonic partials based on a F_0 of 50, 100, or 200 Hz, respectively. The noise maskers were all generated from broadband Gaussian noise. The masker spectral profile was either flat, or speech-shaped, in which case the complex was passed through the same filter designed to match the excitation pattern of average speech. Maskers were not passed through the low- and high-pass filters; only the target bands were. All maskers were 200-ms, with 30-ms onset and offset ramps, and were all equalized at 65 dB SPL. The full design consisted of 6

masker types (3 forms of excitation $\times$ 2 spectral profiles) and 36 conditions in all (6 target bands $\times$ 6 masker types).

B. Model description

The model predicted the audibility of bands of speech-shaped noise in a given masker. It was based on an auditory filter-bank which consisted of 256 rounded-exponential filters, regularly spaced on an equivalent rectangular bandwidth (ERB) scale (Glasberg and Moore, 1990) and realistic phase responses (Oxenham and Dau, 2001). The first part of the model, concerned with spectral masking, was purely based on excitation patterns and therefore disregarded the phase information completely. For each experimental condition, the masker was either generated alone and passed through the filter-bank, or was added to the target at a given TMR and the target + masker compound was passed through the filter-bank. As an example, the top-left panel of Fig. 2 illustrates the excitation pattern of the noise masker with a speech-shaped profile, and that of the compound in which the target is low-pass filtered at 1833 Hz with a TMR of -6 dB (solid lines). Audibility of the target at a given frequency is defined by the difference between the two excitation patterns and appears as a light gray area. This area can be integrated to provide an estimate of how detectable the target is at this TMR. Note that the differences in excitation pattern that occurred below absolute thresholds set at 0 dB hearing level (HL) (represented by the dashed line) were excluded from this integration since none of this energy is supposed to be audible. The top-right panel shows the same target band at the same TMR, in the presence of a RP masker at 200-Hz F_0 , also with a speech-shaped profile. It is apparent that target energy is audible in between resolved partials of the masker, as well as below the fundamental. The integration area for this condition is thus larger than in the noise masker condition, reflecting the predicted benefit of spectral glimpsing.

The second part of the model was a time-analysis that included non-linearity of the BM, which was necessary to predict phase effects. As an example, Fig. 3 shows the temporal outputs for an SP masker (black lines) and a target + SP masker compound (gray lines) in two frequency channels, centered at 1 and 5 kHz. At 5 kHz, the filter is broader than at 1 kHz. In consequence, it passes more partials of the SP masker, resulting in a peakier temporal envelope than at 1 kHz. Within each output, the presence of the faint target set at -20 dB TMR (i.e., 45 dB SPL) in the compound signal is barely visible at the masker dips. The key factor here relates to how this target energy with little intensity obtains differential amplification by the BM. To model this, the Hilbert envelopes were extracted for each output, shown in the second panels of Fig. 3. The dashed line represents absolute threshold at this particular center frequency (set at 0 dB HL) and served as the basis for a threshold device. This threshold device was implemented by half-wave rectification with a direct current offset, so that only the levels above absolute threshold received the non-linear amplification shown in the third panel. This step is critical to reproducing the observed effects, as otherwise the level at the dips of the masker alone would always receive more gain

FIG. 2. Excitation patterns of masker and compound signals in four experimental conditions. The presence of the target is reflected by the gray surface. Comparisons between the two top panels reflect the contribution of spectral dips. Comparisons between the bottom panels reflect the contribution of deep envelope modulations.

than the level at the dips of the compound, thereby reducing instead of enhancing the TMR (Alcántara *et al.*, 1996). Phase effects were thus strongly dependent on the level chosen for absolute thresholds. The fourth panel illustrates the amplified temporal outputs. It is apparent that portions of low intensity receive more amplification than portions of high intensity, making the waveforms less peaky. More importantly, at 1 kHz, the level of the masker alone at the masker dips is not low enough to pass under the threshold device, and consequently both masker alone and compound signals are amplified, so the presence of the target is not enhanced in this case. In contrast, at 5 kHz, absolute threshold divides the level of the compound from the level of the masker alone at the masker dips and as a result, the presence

of the target is enhanced after BM amplification. In order to compare the benefits of BM amplification across center frequencies, and contrast them against the benefits of spectral dips, these benefits had to be translated into excitation levels on a similar scale to that of an excitation pattern. At each center frequency, the root-mean-square (rms) difference was calculated between the masker alone and the compound signal: If this difference had increased after amplification, a difference in excitation level was added. This is illustrated in the dark gray area in the bottom-right panel of Fig. 2, for a target high-passed at 609 Hz with a TMR at -6 dB. Due to the additional MR provided by the non-linear amplification, it is apparent that the total integration area is much larger for a SP speech-shaped harmonic masker at 50-Hz F_0 (bottom

FIG. 3. Description of the time-domain analysis in the model. Stimuli are passed through a bank of 256 auditory filters. The envelope of each within-channel output is extracted and passed through a device representing the threshold for a minimal excitation on the BM. The envelopes are then passed through the input/output function and the temporal outputs are then reconstructed, amplified by the non-linear gain of the BM.

right panel) than for its RP counterpart (bottom left panel), which did not benefit from any additional MR based on non-linear BM amplification in any of the 256 center frequencies.

Note that this implementation of phase effects is not the only possible way. Other studies have used the internal contrast, the difference between the non-linearly amplified envelopes smoothed in time using a temporal integrator (e.g., Alcántara *et al.*, 1996). It is however not clear whether one should extract the maximum or the mean of the internal contrast, tapping into the issue of whether listeners base their decisions on a given instant or over the entire duration of the stimuli. It is also not clear what shape the temporal integrator should have (Oxenham and Moore, 1994), and more particularly what equivalent rectangular duration should it correspond to. Finally, it is not clear how the magnitude of this internal contrast would be translated back into incremental changes in the excitation patterns in order to compare these temporal masking effects with the spectral masking effects. The present modeling has the advantage of making no assumption that listeners actively listen in the dips, no assumption regarding the shape and width of the temporal integrator, and provides estimates of rms difference within each channel of the filter-bank, being therefore comparable with differences in excitation level.

C. Predictions

For each of the 36 experimental conditions, the integration area was calculated at different TMRs between +10 dB and -30 dB, in 2-dB steps (same step size as in the adaptive track), each time with new stimuli. In addition the entire procedure was repeated 17 times (same as the number of subjects), each time with new stimuli. This allowed a fuller representation of target noise bands and masker stimuli. For each “simulated subject,” a masked detection threshold (MDT) may then be predicted for each condition as the lowest TMR at which the integration area is larger than a criterion, fixed across all conditions. A fine resolution in TMR was reached by linearly interpolating the integration area in 0.01-dB steps. The criterion was a fitted parameter, based on the accuracy of predictions for noise maskers. Criteria between 100 and 200 provided reasonable fits to the data obtained for noise maskers: The following predictions originated from a criterion of 180, which led to an r^2 of 0.96. The lines in Figs. 4 and 5 show the mean predictions of MDT for maskers with flat-spectrum profiles and speech-shaped profiles, respectively, with low-pass (left panels) or high-pass (right panels) targets. A number of predictions can be outlined from these modeling results.

Let us first consider the noise masker conditions alone. For a flat-spectrum profile (Fig. 4), the predicted MDTs were much higher for high-pass targets than for low-pass targets and also decreased with low-pass cutoff but increased with high-pass cutoff. These trends can easily be understood from Fig. 1, as they are simply the result of how the distribution of energy in speech compares to the distribution of energy in a flat-spectrum masker. For a speech-shaped profile (Fig. 5), predicted MDTs differed much less between low- and high-pass targets because target and masker had the same spectral

FIG. 4. Mean threshold (symbol) and prediction (line) for the detection of low-pass (left panel) and high-pass (right panel) filtered bands of speech-shaped noise against three flat-spectrum maskers: Sine-phase (white triangles, dashed black lines), random-phase (black circles, solid black lines), and noise maskers (gray squares, interrupted gray lines). Here and later, error bars are ± 1 standard error of the mean over 17 listeners.

shape, but there remained an effect of target bandwidth. The broader the target at a given TMR, the more channels received an increment in excitation level, and therefore the larger the integration area. This explains why predicted MDT still decreased as the low-pass cutoff increased, and why predicted MDT still increased as the high-pass cutoff increased.

More central to this study were the predictions for the RP and SP maskers. The model predicted that (1) MDTs should not differ between the two harmonic maskers in resolved regions, but should differ in unresolved regions, consistent with the idea that phase relationships between harmonic partials only matter in unresolved regions. In resolved regions alone, i.e., for the low-pass targets (left panels of Figs. 4 and 5), the model predicted that (2) MDTs should decrease more with the F_0 of harmonic maskers than what could be predicted from the increase in target bandwidth, consistent with a role for spectral glimpsing between resolved partials. It is notable that this prediction occurs primarily at 200-Hz F_0 , but did not seem to depend strongly on the masker spectral profile. This is more readily illustrated from the dotted lines in the left panel of Fig. 6, representing the difference in predicted MDT between harmonic and noise maskers. In unresolved regions alone, i.e., for the high-pass targets (right panels of Figs. 4 and 5), the model predicted that (3) MDTs should be lower for SP than for RP maskers but that the difference should be reduced as F_0

FIG. 5. Same as Fig. 4 for maskers with a speech-shaped profile.

FIG. 6. (Left) Difference in MDT between noise and harmonic maskers (averaged over RP and SP) in resolved regions. (Right) Difference in MDT between the two harmonic maskers in unresolved regions. Dotted lines represent the model predictions.

increases, consistent with a role for masker envelope modulations within a limited range of F_0 s. In the right panel of Fig. 6, the dotted lines represent the predicted magnitude of these phase effects: They were about 12–15 dB at 50-Hz F_0 , and 9–10 dB at 100-Hz F_0 , and 1.2 dB at 200-Hz F_0 . This trend is very important as it predicts that phase effects disappear as F_0 increases regardless of any additional constraints imposed on the temporal resolution of the differential amplification. Phase effects are lost as F_0 increases simply because the spectral region where partials are largely unresolved is increasingly pushed to high frequencies that overlap less and less with the distribution of energy in speech.

III. EXPERIMENTAL METHOD

A. Listeners

Seventeen listeners took part in the experiment. They were between 20 and 40 yrs old and were paid for their participation. All listeners had pure tone thresholds less than 15 dB HL at frequencies of 0.25, 0.5, 1, 2, 4, 6, and 8 kHz.

B. Stimuli

The same stimuli were used as those described in the computational analysis. Target stimuli were freshly generated for each trial. For the masker stimuli, the same set of random phases or the same noise (but different from that used for the target) was used for all intervals in a trial and freshly generated in each trial. This was intended to limit stimulus uncertainty in the RP and noise masker conditions, which might have produced elevated thresholds compared with the SP masker. The amount of masking caused by stimulus uncertainty is considerably reduced when uncertainty occurs across trials as opposed to within trials (e.g., Neff and Green, 1987).

C. Procedure

Listeners were first familiarized with the task, using one of the 36 conditions at random, for practice. The following 36 runs measured MDT in a 3-interval forced-choice task for each condition in random order. In each trial, the listener heard three intervals: One, selected at random, was the target noise band in the presence of the masker; the other two were the masker alone. The listener was asked to report which

interval contained the target. MDTs were measured using a 2-down/1-up adaptive method. The TMR (defined for the unfiltered target spectrum) was initially 0 dB and varied with 5-dB steps for two reversals and subsequently 2-dB steps for 8 further reversals. MDTs for each run were taken as the mean TMR at the last eight reversals. Subjects were tested only once for each condition. The experiment lasted a little over an hour.

D. Equipment

All experimental data were collected at the Auditory Prostheses and Perception Laboratory at Boys Town National Research Hospital. Signals were sampled at 44.1 kHz and 16-bit resolution, digitally mixed, digital/analog converted by a 24-bit Edirol UA-25 sound card and presented diotically to subjects over Sennheiser HDA 200 headphones in a double-walled sound-attenuating booth manufactured by Industrial Acoustics Company, within a sound-treated room. A computer monitor was inside the booth for trial-by-trial feedback and listeners gave their responses by pressing keys on a keyboard.

IV. RESULTS AND DISCUSSION

The symbols in Figs. 4 and 5 show the mean MDTs (averaged over 17 listeners) obtained with flat-spectrum maskers and speech-shaped maskers, respectively. Statistical analyses first addressed the predictions made in the computational analysis. In a second step, further analyses addressed questions that the model did not capture. Mauchly's test of sphericity was used in all analyses of variance and Greenhouse-Geisser correction was applied to adjust the degree of freedom of one instance where the sphericity assumption was not met.

A. Noise masker conditions

Qualitatively, MDTs were much higher for high-pass targets than for low-pass targets in the flat-spectrum maskers, whereas MDTs were a little lower for high-pass targets than for low-pass targets in the speech-shaped maskers. In both spectral profiles, MDTs varied as expected with the bandwidth of the target band, being lower the broader the band. A repeated-measures analysis of variance was performed on the MDTs for the noise masker alone. The results are reported in Table I. The significant 3-way interaction

TABLE I. MDT for noise maskers.

Factors	<i>F</i> value	<i>p</i> value
profile	$F(1,16) = 581.3$	0.000 *
region	$F(1,16) = 308.6$	0.000 *
cutoff	$F(2,32) = 72.3$	0.000 *
profile $\times$ region	$F(1,16) = 994.8$	0.000 *
profile $\times$ cutoff	$F(2,32) = 78.0$	0.000 *
region $\times$ cutoff	$F(2,32) = 151.9$	0.000 *
3-way	$F(2,32) = 81.7$	0.000 *

was further interrogated by testing the simple effect of cutoff for each factorial combination of profile and region. MDTs decreased as the low-pass cutoff increased in each profile respectively [$F(1,15) = 16.5$ and 10.7 , $p < 0.001$], whereas MDTs increased as the high-pass cutoff increased for the flat-spectrum profile [$F(1,15) = 298.0$, $p < 0.001$], but not for the speech-shaped profile [$F(1,15) = 0.9$, $p = 0.424$]. The profile $\times$ region interaction was also interrogated by testing the simple effect of region for each profile: MDTs were higher for high-pass targets than for low-pass targets in flat-spectrum conditions [$F(1,16) = 943.6$, $p < 0.001$], and conversely in speech-shaped conditions [$F(1,16) = 24.6$, $p < 0.001$]. Predictions were therefore met for this masker type, although there were some offsets with the data which were not trivial to eliminate. For high-pass targets, better predictions could be obtained by taking the maximum of the difference between the excitation patterns of masker and compound signals, instead of integrating this difference across the 256 auditory filters as was done. This measure, on the other hand, would have resulted in much lower predictions for low-pass targets since they displayed large local maxima of MR. These discrepancies may reflect the fact that listeners used different strategies in different frequency regions. The ability to discriminate very local changes in spectral shape might facilitate detection in low-frequency regions. Despite those discrepancies, there was a high correlation between data and predictions, with an r^2 of 0.96. In conclusion, the MDTs for noise maskers were well understood and largely influenced by the effect of masker spectral profile and the effect of target bandwidth.

B. Effect of phase across regions

The key prediction examined here concerned the effect of phase relationships between harmonic partials in resolved or unresolved regions. Potential dependencies on masker profile and F_0 were only of interest within individual region. Thus, for this analysis, MDTs were averaged across these two factors, and a repeated-measures analysis of variance was performed on the mean MDTs for RP and SP maskers in resolved versus unresolved regions. The results are reported in Table II. The significant interaction between phase and region was further interrogated by testing the simple effect of phase in each region. There was no systematic effect of phase in resolved regions but a clear phase effect in unresolved regions. The data therefore confirmed that the phase relationships between harmonic partials only mattered in unresolved regions.

TABLE II. MDT for RP and SP maskers.

Factors	<i>F</i> value	<i>p</i> value
phase	$F(1,16) = 31.6$	0.000 *
region	$F(1,16) = 154.7$	0.000 *
phase $\times$ region	$F(1,16) = 88.3$	0.000 *
phase in resolv. region	$F(1,16) = 0.6$	0.435
phase in unresolv. region	$F(1,16) = 79.0$	0.000 *

C. Resolved regions alone

Since there was no effect of phase in the resolved regions, the MDTs were averaged over RP and SP maskers and the difference in MDT between noise and harmonic maskers was calculated, shown by the bars in the left panel of Fig. 6. The key prediction examined here was that this difference should increase with F_0 because there were more and more opportunities to glimpse target energy in between resolved masker partials. A repeated-measures analysis of variance was performed on these differences in MDT. The results are reported in Table III. The main effect of F_0 reflected that the differences increased by 5 to 6 dB between 50- and 200-Hz F_0 s, but did not interact with masker spectral profile. As suggested by the model, the size of spectral dips differs little whether resolved partials belong to a flat-spectrum or a speech-shaped masker. Consequently, the MR provided by spectral glimpsing does not depend much on the masker spectral profile.

It is however notable that the difference in MDT between noise and harmonic maskers increased particularly between 50- and 100-Hz F_0 in the data (especially for the flat-spectrum) whereas it increased particularly between 100- and 200-Hz F_0 in the prediction. This discrepancy as well as the large underestimation of the size of these differences may be due to a mechanism not implemented in the model (Sec. IV E).

D. Unresolved regions alone

Apart from the fact that phase effects did occur in unresolved regions (Sec. IV B), the key prediction examined here was that these phase effects should progressively disappear as F_0 increased. The difference in MDT between RP and SP maskers in unresolved regions was calculated and shown by the bars in the right panel of Fig. 6. A repeated-measures analysis of variance was performed on these differences in MDT. The results are reported in Table IV. There was a main effect of F_0 reflecting the progressive reduction in phase effect as F_0 increased, and it also interacted with masker spectral profile since this reduction was more dramatic with speech-

TABLE III. Difference in MDT between the harmonic maskers (averaged over RP and SP) and noise maskers in resolved regions.

Factors	<i>F</i> value	<i>p</i> value
profile	$F(1,16) = 36.4$	0.000 *
F_0	$F(1.2,19.5) = 7.0$	0.012 *
profile $\times$ F_0	$F(2,32) = 1.7$	0.200

TABLE IV. Difference in MDT between RP and SP maskers in unresolved regions.

Factors	<i>F</i> value	<i>p</i> value
profile	$F(1,16) = 1.4$	0.247
<i>F</i> 0	$F(2,32) = 125.2$	0.000 *
profile $\times$ <i>F</i> 0	$F(2,32) = 34.9$	0.000 *

shaped than with flat-spectrum maskers. The magnitude of these phase effects was further interrogated for each factorial combination of *F*0 and profile. Phase effects were very large at 50 Hz for both flat-spectrum [$F(1,16) = 43.1$, $p = 0.000$] and speech-shaped [$F(1,16) = 130.4$, $p = 0.000$] spectral profiles, with magnitudes of 10.5 and 18.3 dB, respectively. Phase effects were also significant at 100 Hz for both flat-spectrum [$F(1,16) = 21.0$, $p = 0.000$] and speech-shaped [$F(1,16) = 5.1$, $p = 0.038$] spectral profiles, with magnitudes of 7.3 and 3.6 dB, respectively. The phase effect was however not significant for flat-spectrum maskers at 200-Hz *F*0 [$F(1,16) = 0.8$, $p = 0.387$], and was significant in the opposite direction for speech-shaped maskers at 200-Hz *F*0 [$F(1,16) = 4.7$, $p = 0.045$], i.e., mean MDT was 3.5 dB lower for RP than SP. Masker envelope modulations were thus progressively less useful as *F*0 increased, but the dependency of phase effects on *F*0 was stronger for speech-shaped maskers than for flat-spectrum maskers.

E. Effects not captured by the model

The experimental data confirmed that (1) spectral dips in harmonic maskers provide some MR in resolved regions, which increases with *F*0; and (2) that envelope modulations resulting from the interactions of unresolved partials with specific phase relationship, such as sine phase, provide some MR which decreases with *F*0. Due to these opposite trends, energy of a speech-shaped source is more likely to be detectable in high center frequencies against harmonic maskers with low *F*0s, while it is more likely to be detectable in low center frequencies against harmonic maskers with high *F*0s. In addition, there were important aspects of the data which were ignored by the model and have the potential to change the balance between target information in resolved and unresolved regions.

1. Masker periodicity

The most striking flaw of the model concerns the difference in MDT between harmonic and noise maskers. For instance, the model predicted that spectral dips were too shallow in harmonic complexes with a *F*0 of 50 Hz to provide any MR compared to noise maskers in resolved regions. In contrast, the experimental data showed between 6- and 12-dB MRs at this *F*0 for flat-spectrum and speech-shaped profiles, respectively (left panel of Fig. 6). This limitation is also clearly apparent in unresolved regions, where the model predicted identical thresholds for RP and noise maskers, since there were no spectral dips at these high center frequencies, whereas the experimental data showed between 5- and 10-dB MRs for the RP masker. One explanation for

FIG. 7. MR provided by the use of masker periodicity, as a function of center frequency, taken from Deroche and Culling (2011).

these discrepancies is that masker periodicity in within-channel fine structures also provides some MR which was not implemented here.

Deroche and Culling (2011) found evidence for such a mechanism by measuring MDTs for 100-Hz wide noise bands, with different center frequencies, masked by a flat-spectrum harmonic or inharmonic complex based on a 100-Hz *F*0 reference. One masker partial centered in the middle of the target band was fixed in frequency across the harmonic and inharmonic cases to ensure that the excitation level was identical over the target bandwidth, thereby preventing differences in spectral dips of the complexes from affecting the MDT. In addition, both complexes had partials in random phase to prevent deep envelope modulations. They observed MRs for center frequencies between 0.5 and 2.5 kHz, reproduced in Fig. 7. More importantly these MRs were partly based upon periodicity in remote frequency channels, suggesting an across-channel integration of periodicity. Our current understanding of this mechanism remains nonetheless relatively poor. For instance, it is not yet clear how these MRs would vary with TMR in the case of a speech-shaped target. This uncertainty prevents a quantitative analysis. From a qualitative point of view nonetheless, this distribution of MR with center frequency (Fig. 7) suggests that the use of periodicity may benefit information in low rather than in high center frequencies. Since speech-shaped maskers have relatively more energy in low than in high center frequencies, it may well be that masker periodicity provides more MR with a speech-shaped masker than with a flat-spectrum masker, since there is more masking to begin with. The role of masker periodicity could thus explain why the difference in MDT between harmonic and noise maskers was larger with a speech-shaped profile than with a flat-spectrum profile (left panel of Fig. 6).

MDTs for the noise maskers were still higher than for the RP maskers in high-frequency regions where Deroche and Culling's data suggest that periodicity has little impact. For instance, Fig. 7 shows very little benefit of periodicity above 3122 Hz (25 ERBn), and yet there were 7- to 10-dB differences in MDT between noise and RP maskers at

200-Hz F_0 . So it is likely that a fourth mechanism, modulation masking, is also at play.

2. Modulation masking

Listeners may be able to detect the presence of a target noise band by the fluctuations it produces, given a reasonable window of temporal integration, in comparison with the fluctuations of the masker alone. Figure 8 illustrates this account for three speech-shaped maskers, for a target high-pass filtered above 3122 Hz at a TMR of -10 dB. The sliding temporal integrator was based on a rounded exponential window with the shortest equivalent rectangular duration provided by Oxenham and Moore (1994; Table I). In the left panel, the envelope level of the noise masker fluctuates substantially and in a random way, making detection of the target noise on the basis of random fluctuations extremely difficult. In contrast, in the middle panel, the envelope level of the RP masker is quite flat, which makes it easier to detect the target-induced random fluctuations. This difference would explain why MDTs were lower for RP maskers than for noise maskers in unresolved regions, an effect not predicted by the model. Furthermore, the higher the F_0 , the flatter the outputs of the temporal integrator for the RP masker, which could also explain why MDTs differed more and more between noise and RP maskers as F_0 increased in unresolved regions.

Another intriguing aspect of the data is the reversal in phase effect observed at 200-Hz F_0 with the speech-shaped spectral profile in unresolved regions: MDT was on average 3.5 dB lower for RP than SP masker (black circle and white triangle in the right panel of Fig. 5). This reversed phase effect had also been observed by Gockel *et al.* (2002) for the detection of a target noise masked by complexes based on 250-Hz F_0 , bandpass filtered in the region of unresolved partials. MDT was about 2 dB lower for the RP masker than for a cosine phase masker. They also pointed out the difference between the within-channel envelopes of RP and SP maskers. The waveform of the SP masker is a lot peakier than that of the RP masker due to the specific phase relationship between partials passing through this filter. These modulations are smoothed to some extent by the temporal integrator, but there remain some periodic fluctuations at the rate of F_0 as shown in the right panel of Fig. 8. These

fluctuations may hinder the listener's ability to detect the target-induced fluctuations. This reversal in phase effects is very important as it suggests that strong envelope modulations in a masker waveform are not always beneficial. Furthermore, effects of modulation masking may not be restricted to F_0 s of 200 Hz or higher: At 100-Hz F_0 , the predicted phase effects were over-estimated compared to the phase effects observed experimentally, especially for the speech-shaped profile. If such modulation masking effects can occur for F_0 s as low as 100 Hz, this is a very serious issue to consider since they would largely overlap with F_0 s in the human voice range.

F. Is speech more likely to be recovered from resolved or unresolved regions of harmonic maskers?

Having gained insights into the different mechanisms involved in situations of masking by harmonic backgrounds, it is now important to compare MDT for a given complex between the two regions, in order to conclude on the frequency region where energy of a speech-shaped source is most likely to be available. Paired-samples t tests were performed between the two regions for each harmonic masker, each spectral profile, and each F_0 , with Bonferroni corrections, i.e., at a significance level of 0.05/12. The results are reported in Table V. For partials in random phase, MDTs were always lower in resolved than in unresolved regions, except at 50-Hz F_0 with the speech-shaped profile where MDT was similar in both regions. For partials in sine phase, MDTs were lower in unresolved than in resolved regions at 50-Hz F_0 , and vice versa at 200-Hz F_0 . At 100-Hz F_0 , MDTs were either similar or lower in resolved than in unresolved regions.

Phase effects in MDT could transfer largely to speech recognition with a masker F_0 at 50 Hz, not so much because the masker envelope level falls in the compressive part of the input-output function, but rather because almost the entire speech spectrum falls in the unresolved region. For masker F_0 s of 100 Hz and above, speech energy is present in both resolved and unresolved regions, but is generally audible at a lower TMR in resolved regions than in unresolved regions. This relatively low-frequency information may be sufficient to obtain 50% intelligibility and therefore masker envelope modulations may not necessarily contribute to

FIG. 8. Illustration of a phenomenon related to modulation masking. Listeners may detect a target band of noise by the random fluctuations it produces in comparison with a background fluctuation. After temporal integration, envelopes of the RP masker are relatively flat and facilitate detection of these random fluctuations. In contrast, the envelopes of the SP masker, and to a greater extent those of the noise masker, hinder detection of the target-induced random fluctuations.

TABLE V. Paired-samples t tests revealed the conditions in which MDTs were lower in resolved than in unresolved regions, marked as *, and the conditions in which MDTs were lower in unresolved than in resolved regions marked as #.

	RP masker		SP masker	
	Flat-spectrum	Speech-shaped	Flat-spectrum	Speech-shaped
50 Hz	$t(16) = -3.8, p = 0.0016 *$	$t(16) = -0.5, p = 0.6080$	$t(16) = 5.9, p < 0.0001 \#$	$t(16) = 18.2, p < 0.0001 \#$
100 Hz	$t(16) = -21.1, p < 0.0001 *$	$t(16) = -8.7, p < 0.0001 *$	$t(16) = -7.4, p < 0.0001 *$	$t(16) = 0.2, p = 0.8567$
200 Hz	$t(16) = -12.6, p < 0.0001 *$	$t(16) = -3.4, p = 0.0035 *$	$t(16) = -10.1, p < 0.0001 *$	$t(16) = -8.8, p < 0.0001 *$

speech recognition. This is essentially what [Deroche et al. \(2013\)](#) observed when measuring SRT for an intonated target voice, male or female, against harmonic complexes with either flat or speech-shaped spectral profiles, at 50-, 100-, or 200-Hz F_0 and presented at a fixed level of 65 dB SPL. Their results are reproduced in Fig. 9, averaged over talker gender which had little influence. SRT for RP maskers progressively decreased as masker F_0 increased, reflecting the progressive role of spectral glimpsing. In contrast, SRT for a SP masker was 10 dB lower than for its RP counterpart at 50-Hz F_0 , due to the large MR based upon masker envelope modulations, while this benefit was lost for F_0 s of 100 Hz and above. As a conclusion, the key factors to speech recognition in the presence of harmonic maskers are the TMRs at which different speech cues become audible and how essential these cues are for intelligibility. If enough speech information can be recovered from the region of resolved masker partials exclusively, then it may not matter that the specific phase relationships between masker partials result in strong envelope modulations in unresolved regions.

V. SUMMARY

In the presence of harmonic maskers, there are spectral dips available in resolved regions, regardless of the phase relationships between partials, and there are temporal dips available in unresolved regions, depending on the phase relationships between partials. As the F_0 of a complex masker increases, there are more and more opportunities for spectral glimpsing and less and less opportunities for the BM to differentially amplify target energy at temporal dips in the masker waveforms. These two mechanisms are sufficiently well understood to attempt to model them. The present implementation

listed a number of key predictions, which were observed in the experimental data. First, MDTs did not differ systematically with the phase relationships between harmonic partials in resolved regions, but decreased with the masker F_0 , and particularly more than what could be predicted from the increase in target bandwidth, consistent with the benefit of spectral glimpsing. Furthermore, the effect of masker F_0 was similar with a flat-spectrum as with a speech-shaped masker. Second, MDTs were lower for SP than RP maskers in unresolved regions, confirming that listeners could also benefit from the deep envelope modulations of the SP masker at high center frequencies. However, these phase effects disappeared rather quickly as F_0 increased, because the region of unresolved partials was pushed to very high center frequencies overlapping less and less with the distribution of speech energy. Finally, the magnitude of phase effects and their dependency on F_0 was dependent on the masker spectral profile. Phase effects were larger at 50-Hz F_0 with the speech-shaped than with the flat-spectrum maskers, but decreased more dramatically as F_0 increased with speech-shaped than with flat-spectrum maskers. Therefore, there is sort of a competition between the MRs provided from spectral dips and the MRs provided from envelope modulations, in that there may be sufficient information about a speech source in one region to not need information from the other, and the balance between the two regions depends strongly on the masker F_0 and to a smaller extent on the spectral profile.

There were important aspects of the data that were not predicted by the model which also play an active role in this competition. First, there were large differences in MDT between harmonic and noise maskers in both resolved and unresolved regions. It seems plausible that periodicity in the fine structure provides some additional MR, particularly in frequency regions below about 3 kHz. Second, the large MR between noise and harmonic maskers at very high center frequencies as well as the reversal in phase effect at 200-Hz F_0 in unresolved regions suggested that some additional MR may come from a mechanism that detects target-induced random fluctuations in envelope level from a flatter background, a form of modulation MR. Without a better understanding of these additional mechanisms, it is difficult to reach reliable predictions of the audibility of speech energy across the spectrum. Final comparisons between the measured MDTs in the two regions suggest that unless masker F_0 s are very low, such as 50 Hz, energy in a speech-shaped source is generally available at lower TMR in resolved regions than in unresolved regions. This seems to be because in addition to the MR provided by spectral glimpsing, the masker periodicity may provide more MR in relatively low frequency regions, and

FIG. 9. SRTs collected by [Deroche et al. \(2013\)](#) for a voice masked by SP (white circles) or RP (black circles) complexes with flat or speech-shaped spectral profiles.

strong envelope modulations in high frequency regions may be detrimental rather than beneficial at high F_0 s.

ACKNOWLEDGMENTS

This work was supported by NIH Grant Nos. R01 DC004786, R01 DC004786-08S1, and R21 DC011905 to M.C., and the Human Subjects Research Core of NIH Grant No. P30 DC004662 to Michael Gorga.

- Alcántara, J. I., Holube, I., and Moore, B. C. J. (1996). "Effects of phase and level on vowel identification: Data and predictions based on a nonlinear basilar-membrane model," *J. Acoust. Soc. Am.* **100**, 2382–2392.
- Alcántara, J. I., Moore, B. C. J., Glasberg, B. R., Wilkinson, A. J. K., and Joras, U. (2003). "Phase effects in masking: Within- versus across-channel processes," *J. Acoust. Soc. Am.* **114**, 2158–2166.
- ANSI (1997). ANSI S3.5-1997. *Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York, 1997).
- 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.
- Carlyon, R. P., and Datta, A. J. (1997). "Excitation produced by Schroeder-phase complexes: Evidence for fast-acting compression in the auditory system," *J. Acoust. Soc. Am.* **101**, 3636–3647.
- de Cheveigné, A. (1997). "Concurrent vowel segregation. III. A neural model of harmonic interference cancellation," *J. Acoust. Soc. Am.* **101**, 2857–2865.
- 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.
- Deroche, M. L. D., and Culling, J. F. (2011). "Narrow noise band detection in a complex masker. Masking level difference due to harmonicity," *Hear. Res.* **282**, 225–235.
- Deroche, M. L. D., Culling, J. F., and Chatterjee, M. (2013). "Phase effects in masking by harmonic complexes: Speech recognition," *Hear. Res.* **306**, 54–62.
- Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.
- Gockel, H., Moore, B. C. J., and Patterson, R. D. (2002). "Asymmetry of masking between complex tones and noise: The role of temporal structure and peripheral compression," *J. Acoust. Soc. Am.* **111**, 2759–2770.
- Green, T., and Rosen, S. (2013). "Phase effects on the masking of speech by harmonic complexes: Variations with level," *J. Acoust. Soc. Am.* **134**, 2876–2883.
- Kohlrausch, A., and Sander, A. (1995). "Phase effects in masking related to dispersion in the inner ear. II. Masking period patterns of short targets," *J. Acoust. Soc. Am.* **97**, 1817–1829.
- Meddis, R., and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.* **91**, 233–245.
- Neff, D. L., and Green, D. M. (1987). "Masking produced by spectral uncertainty with multicomponent maskers," *Percept. Psychophys.* **41**, 409–415.
- Oxenham, A. J., and Dau, T. (2001). "Towards a measure of auditory-filter phase response," *J. Acoust. Soc. Am.* **110**, 3169–3178.
- Oxenham, A. J., and Moore, B. C. J. (1994). "Modeling the additivity of non-simultaneous masking," *Hear. Res.* **80**, 105–118.
- 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.
- Shackleton, T. M., and Carlyon, R. P. (1994). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation," *J. Acoust. Soc. Am.* **95**, 3529–3540.
- Summers, V., and Leek, M. R. (1998). "Masking of tones and speech by Schroeder-phase harmonic complexes in normally hearing and hearing-impaired listeners," *Hear. Res.* **118**, 139–150.
- Yates, G. K. (1990). "Basilar membrane nonlinearity and its influence on auditory nerve rate-intensity functions," *Hear. Res.* **50**, 145–162.