Published in final edited form as:
Hear Res. 2013 December ; 306: . doi:10.1016/j.heares.2013.09.008.

Phase effects in masking by harmonic complexes: Speech recognition

Mickael L. D. Deroche[1], **John F. Culling**[2], and **Monita Chatterjee**[3]

[1]Department of Otolaryngology, Johns Hopkins University School of Medicine, 818 Ross Research Building, 720 Rutland Avenue, Baltimore, MD 21205.

[2]School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT.

[3]Auditory Prostheses and Perception Laboratory, Boys Town National Research Hospital, 555 N 30th Street, Omaha, NE 68131.

Abstract

Harmonic complexes that generate highly modulated temporal envelopes on the basilar membrane (BM) mask a tone less effectively than complexes that generate relatively flat temporal envelopes, because the non-linear active gain of the BM selectively amplifies a low-level tone in the dips of a modulated masker envelope. The present study examines a similar effect in speech recognition. Speech reception thresholds (SRTs) were measured for a voice masked by harmonic complexes with partials in sine phase (SP) or in random phase (RP). The masker's fundamental frequency (F0) was 50, 100 or 200 Hz. SRTs were considerably lower for SP than for RP maskers at 50-Hz F0, but the two converged at 100-Hz F0, while at 200-Hz F0, SRTs were a little higher for SP than RP maskers. The results were similar whether the target voice was male or female and whether the masker's spectral profile was flat or speech-shaped. Although listening in the masker dips has been shown to play a large role for artificial stimuli such as Schroeder-phase complexes at high levels, it contributes weakly to speech recognition in the presence of harmonic maskers with different crest factors at more moderate sound levels (65 dB SPL).

1. Introduction

Tone detection in noise and speech recognition in noise both improve when the noise is temporally modulated rather than steady for normal-hearing listeners (de Laat and Plomp, 1983; Festen and Plomp, 1990). Because the target-to-masker ratio (TMR) is higher during the low-level portions of the masker than during its high-level portions, the target is easier to detect at certain times over the masker duration. Listeners seem able to use these dips in the broadband temporal envelope of a masker to detect the target, an ability often referred to as "listening in the dips" or "temporal glimpsing". More impressive, perhaps, is the finding that the auditory system gains similar benefits when detecting pure tones during the extremely short temporal dips that occur over the period of a harmonic complex (Kohlrausch and Sander, 1995; Carlyon and Datta, 1997b). This phenomenon is responsible for variations in the masking potency of harmonic complexes with different phase spectra, and, consequently

[1]Author to whom correspondence should be addressed, Electronic mderoch2@jhmi.edu.

different modulation envelopes. Several aspects of the auditory periphery are thought to contribute to these effects.

1.1. Basilar membrane compression

Outer hair cells amplify an input signal non-linearly as a function of its level (Johnstone *et al.*, 1986; Yates, 1990). As shown in the input/output characteristic for a fixed place along the BM (Ruggero *et al.*, 1997), there is a linear amplification for input levels below 20 dB SPL, which is progressively compressed as level increases up to about 80 dB SPL, beyond which there is no amplification. Thus for a fluctuating signal, low-level portions are enhanced relative to high-level portions. Sine- or cosine-phase harmonic complexes present highly modulated temporal envelopes along the BM: the peaks are compressed relative to the dips. In contrast, random-phase harmonic complexes tend to produce relatively flat envelopes along the BM: little differential gain occurs within a period. Alcántara *et al.* (2003) showed that BM compression could account for the masking potency of cosine- and random-phase complexes: within an individual filter, detection of a pure tone was easier for maskers with peaky envelopes than for maskers with flat envelopes. Many studies have shown similar masking differences, but used Schroeder-phase complexes instead (Schroeder, 1970); for these stimuli, masking differences do not depend solely on BM compression but also on the phase curvature of auditory filters.

1.2. Compression combined with phase curvature within individual filters

Several studies (Smith *et al.*, 1986; Kohlrausch and Sander, 1995; Carlyon and Datta, 1997a, 1997b; Summers and Leek, 1998) have reported that masked detection of a pure tone was easier for positive than for negative Schroeder-phase complexes. These complexes both have flat temporal envelopes, but their phase curvature is such that once they have passed through the phase response of individual auditory filters, at for instance 1 kHz, positive Schroeder-phase complexes produce peaky within-channel envelopes whereas negative Schroeder-phase complexes produce flat envelopes. The two types consequently receive differential gain in the cochlea. Thus, the phase effects observed in these studies can be attributed to BM compression combined with the phase curvature within individual filters.

The phase response of individual auditory filters has been investigated both physiologically and psychoacoustically. Physiological measurements of BM vibration (de Boer and Nuttall, 1997; Recio *et al.*, 1998) and auditory nerve response (Carney *et al.*, 1999) reveal frequency glides in the impulse responses, suggesting that the phase response at a given place on the BM has a negative curvature with frequency. Psychoacoustic estimations of phase curvature used stimuli with different phase spectra. These spectra were based on the modified Schroeder equation suggested by Lentz and Leek (2001). The resulting complexes have a phase curvature coefficient C, which can either counteract ($C>0$) or accentuate ($C<0$) the phase dispersal of the BM response, resulting in modulated or flat internal envelopes, respectively. The C value for which the complex is least efficient at masking a pure tone provides an indirect measure of the phase curvature of an individual filter centered at the pure tone frequency. Several studies (Oxenham and Dau, 2001b, 2004; Lentz and Leek, 2001; Shen and Lentz, 2009) thus derived estimates of phase curvature which were negative in sign but closer to 0 at high center frequencies.

A recent study (Wojtczak and Oxenham, 2009) showed that the combination of compression and negative curvature was in some cases insufficient to explain phase effects observed in forward masking conditions. They suggested a more complex picture in which the medial olivo-cochlear reflex could modulate the non-linearity of the BM depending on the spectral region and the masker duration. The outcome of such complex interactions for speech recognition is currently unknown.

1.3. Phase effects in speech recognition

Most of the literature on phase effects in masking examines detection of pure tones and focuses often on the curvature at a single place on the BM. The question arises as to whether similar phase effects can be observed for processing of broadband signals such as in speech recognition. Speech intelligibility requires integration of information across frequency (Fletcher and Galt, 1950). Phase effects might be observed when masker internal envelopes are deeply modulated at a range of places on the BM. Any Schroeder complex with a positive curvature coefficient may counteract the negative curvature over a range of auditory filters and provided that this range is sufficiently broad, there may be detectable improvements in speech understanding owing to compression. Summers and Leek (1998) investigated the recognition of a female voice in the presence of Schroeder-phase complexes based on a 100-Hz F0. They found 8-10 dB difference in speech reception threshold (SRT) between positive and negative Schroeder-phase maskers. This phase effect was not observed in hearing-impaired listeners. They concluded that the phase effects they observed for tone detection transferred to speech recognition: for listeners with healthy cochleae, BM compression facilitated detection of speech information at dips in deeply modulated internal masker envelopes.

Although Schroeder-phase harmonic complexes are a very valuable tool for investigations of phase sensitivity in auditory perception, they are arguably not the most ecological stimuli to examine. In a typical cocktail-party situation (Hawley *et al.*, 2004), a target voice is masked by other voices. At a given place on the BM, partials of the human voice are roughly in phase, which makes sine-phase complexes ecologically relevant, at least for nearby masking voices. In a room, however, the combined reflections from the walls produce random modifications to the partials' phases: random-phase complexes are consequently ecologically relevant for listening in reverberant conditions. Moreover, compared with flat-spectrum maskers (such as Schroeder-phase complexes), a masking voice has relatively more energy in low frequencies and less energy in high frequencies. This speech-shaped spectral profile might reduce phase effects, because they occur primarily in high-frequency regions where partials are unresolved. In line with our objective of assessing real world relevance, the present study re-examined whether phase effects transfer to speech recognition, a) for sine- and random-phase harmonic maskers, b) for different F0s, 50, 100 and 200 Hz, c) for different spectral profiles, flat-spectrum or speech-shaped maskers, and d) for different talker genders.

2. Method

2.1. Listeners

Two groups of twelve listeners took part in the experiment. They were aged between 20 and 40 years old and were paid for their participation. All listeners had pure tone thresholds less than 15 dB HL at frequencies between 0.25 and 8 kHz and English was their first language. None of them were familiar with the sentences used during the test. All listeners were given two practice blocks, prior to data collection. Each of the two experimental sessions lasted 30 to 45 minutes.

2.2. Stimuli and conditions

One group of listeners heard a male voice, while the other group heard a female voice. Both male and female voices were unprocessed and normally intonated. The corpus of sentences was taken from the Harvard Sentence List (IEEE, 1969). The sentences have low predictability and each has five keywords highlighted in capitals. For instance, one target sentence was "the YOUNG GIRL GAVE no CLEAR RESPONSE." Two sets of sixty sentences were used, i.e. a total of 120 sentences.

There were two types of maskers, sine-phase (SP) and random-phase (RP) harmonic complexes that consisted of 440, 220, or 110 harmonic partials based on a F0 of 50, 100, or 200 Hz respectively. They were either flat-spectrum in which case all partials had equal amplitude, or speech-shaped in which case they were passed through a linear-phase FIR filter designed to match the excitation pattern of average speech. The average speech was based on a larger corpus of 450 sentences spoken by the male talker and the same 450 sentences spoken by the female talker. Thus in the speech-shaped masker conditions, the two groups heard different talkers (male or female) but the same gender-averaged speech-shaped maskers.

In the flat-spectrum maskers, many partials were located in high spectral regions where there was little or no speech information. Nonetheless several studies have shown that partials remote from a given center frequency can still contribute to lowering masked threshold at that frequency (Carlyon and Datta, 1997b; Oxenham and Dau, 2001a; Alcantara *et al.*, 2003). Given the asymmetric shape of auditory filters, low-frequency partials may indeed enhance the envelope modulations in high center frequencies. It is much less likely that high-frequency partials could enhance the envelope modulations in low center frequencies, but a mechanism akin to comodulation masking release (CMR) may occur. Correlated envelope fluctuations at high center frequencies could cue listeners about the presence of a tone in lower center frequencies. Although across-channel comparisons could facilitate detection of pure tones, it is much less clear how substantial the CMR would be with broadband complex stimuli such as speech. It was therefore interesting to examine whether the presence of high-frequency masker partials could influence the ability to listen in the masker dips, as indexed by the size of observed phase effects.

The RP complexes were frozen, i.e., a fixed set of random phases was generated for each F0 and used throughout the experiment. To avoid the possibility that an unfortunate set of random phase would lead to well-modulated envelopes, many versions of RP complexes were generated. A detailed analysis of crest factors at the output of a bank of simulated auditory filters was performed for each version until one version was found to display particularly flat within-channel temporal envelopes (see Appendix for a complete description). This modeling work was similar to that performed by Hartmann and Pumplin (1991) but with many more partials, for which it becomes very difficult to find the single waveform with minimal fluctuations. The resulting phase structures of these selected RP complexes displayed no obvious order, and remained very dissimilar to Schroeder-phase complexes. Within many individual filters, phase values varied erratically. As shown in Figure A2, crest factors varied by as much as one unit across different center frequencies, as well as across different RP waveforms. On average, however, over all 256 filters, these complexes had flatter temporal envelopes than most RP complexes would have. As an example, Figure 1 compares the temporal signals of the SP and RP flat-spectrum complexes, based on a 50-Hz F0, at the output of simulated filters centered at 503, 1989, and 5004 Hz. At all three center frequencies, the temporal envelopes of the RP complex (right panels) are weakly modulated at a rate of 50 Hz. In contrast, the temporal envelopes of the SP masker (left panels) are modulated with greater depth and the effect is exacerbated as center frequency increases, consistent with broader filters at more basal regions of the BM.

Sentences spoken by the male voice ranged in duration between 1.9 and 4.0 seconds, with a mean of 2.6 seconds. Sentences spoken by the female voice ranged in duration between 2.7 and 4.4 seconds, with a mean of 3.6 seconds. The maskers were always 4.5 seconds long; they were not gated on and off together with the target sentences. All maskers and target sentences were all equalized to the same RMS power, prior to level changes during the adaptive track. Maskers were presented at 65 dB SPL and the relative target level was adjusted.

2.3. Procedure

The first experimental session began with two practice runs using broadband Gaussian white noise as a masker, in order to familiarize listeners with the task. The following six runs measured SRT in the presence of flat-spectrum maskers. The second experimental session occurred immediately after the first one without practice: six runs measured SRTs in the presence of speech-shaped maskers. The two sessions used two different sets of 60 sentences, so no sentence was presented twice to a participant. Effects of order and materials could only occur for the masker profile factor (flat-spectrum/speech-shaped). For the three other factors (F0, phase, and talker gender), effects of order and materials were counterbalanced within each experimental session: while each of the 60 target sentences was presented to every listener in the same order, the order of the conditions was rotated for successive listeners. Twelve listeners (in each group) resulted in two complete rotations of the conditions.

SRT was measured using a 1-up/1-down adaptive method (Plomp and Mimpen, 1979). In this method, ten target sentences are presented one after another, each one against the same masker. The TMR is initially very low (−32 dB) and in the initial phase, listeners have the opportunity to listen to the first sentence a number of times, each time with a 4-dB increase in TMR. Listeners are instructed to attempt to type a transcript of the first sentence when they believe that they can first hear about half the words of the target sentence. The correct transcript is then displayed on the screen, with five key words in capitals, and the listener self-marks how many key words he/she gets correct. Subsequent target sentences are presented only once and self-marked in a similar manner; the level of the target speech is decreased by 2 dB if the listener correctly identifies three or more of the five key words or else increased by 2 dB. Measurement of each SRT is taken as the mean TMR at the last eight trials.

2.4. Equipment

A computer monitor was inside the booth for trial-by-trial feedback and listeners gave their responses by typing their transcriptions on a keyboard. Signals were sampled at 44.1 kHz and 16 bits, digitally mixed, D/A converted by a 24-bit Edirol UA-25 sound card and presented diotically to subjects over Sennheiser HDA 200 headphones in a double-walled IAC sound-attenuating booth within a sound-treated room.

3. Results

Figure 2 presents the mean SRTs measured with flat-spectrum (left panel) or speech-shaped (right panel) maskers. A four-factor analysis of variance with three within-subjects factors (F0 × phase × masker profile) and one between-subjects factor (talker gender) was conducted in order to determine the influence of each factor on mean SRT. The statistics are reported in Table I.

The main effect of masker profile reflected the fact that, on average, SRTs were 5.0 dB higher for speech-shaped than for flat-spectrum maskers. As expected, speech-shaped maskers were more effective maskers since they matched the excitation pattern of a typical voice. The main effect of talker gender reflected the fact that, on average, SRTs were 2.1 dB lower for the female voice than the male voice; however there was a strong interaction between masker profile and talker gender. Post-hoc pairwise comparisons indicated that SRTs were not different for the two voices when maskers were flat-spectrum [$F(1,22)=0.2$, $p>0.05$], but were 4.0 dB lower for the female than the male voice when the maskers were speech-shaped [$F(1,22)=37.5$, $p<0.001$]. In other words, any argument that the female voice was intrinsically more intelligible than the male voice is not supported, because SRTs did

not depend on gender in the flat-spectrum conditions. The effect of talker gender reflected the fact that the gender-averaged speech-shaped masker was less effective in masking the female voice than the male voice. This effect can presumably be accounted for by differences in average spectra between the male and female voice. On average over the 120 target sentences, the excitation level above 3 kHz could be up to 8 dB higher for the female than the male voice.

There was a main effect of F0 as well as a main effect of phase, but more interesting was the strong interaction between F0 and phase, which did not depend on the talker gender and did not depend on whether the maskers were flat-spectrum or speech-shaped. Post-hoc pairwise comparisons indicated that SRTs were on average 10.3 dB lower for the SP than for the RP masker at 50-Hz F0 [$F(1,22)=211.1$, $p<0.001$]. In contrast, SRTs were not different between the SP and RP maskers at 100-Hz F0 [$F(1,22)=1.2$, $p>0.05$]. Finally, SRTs were 2.2 dB higher for the SP than for the RP maskers at 200-Hz F0 [$F(1,22)=23.6$, $p<0.001$]. These differences are discussed below at each F0 separately.

Finally, phase and masker spectral profile interacted and this interaction was stronger for the female voice than for the male voice, resulting in a 3-way interaction. These interactions did not, however, involve F0, so their interpretation is not trivial given that the size and even the sign of phase effects changed across the three F0s. As discussed below, different mechanisms may be involved at different masker F0s, so any interpretation of these interactions would be speculative.

4. Discussion

4.1. 50-Hz F0

At 50-Hz F0, the present results showed that speech recognition was better when it was masked by a SP harmonic complex than when it was masked by a RP complex. This masking release, about 10.3 dB, is referred to as the *phase effect*. At the output of several broad auditory filters, the envelope is strongly modulated when partials interact in phase (left panels of Fig. 1), but weakly modulated when partials interact randomly (right panels of Fig. 1). Listeners could therefore have glimpsed speech information at dips in the SP masker waveform, but not in the flatter RP masker waveforms.

Observing a phase effect, however, does not guarantee the involvement of cochlear compression. Listening in temporal dips as long as 20 ms might occur without speech being selectively amplified at dips in the masker waveform. In the literature, evidence for compression arises from the level dependency of these phase effects. In the present study, masker level was fixed at 65 dB SPL, and it is not obvious how the phase effect observed at 50-Hz F0 would vary with masker level. Amplitude peaks in the within-channel envelope of the SP maskers had levels between 45 and 70 dB (for center frequencies up to 10 kHz), i.e., within the compressive range of the input-output function. It is thus at least possible that speech information located at dips in the SP masker envelopes was selectively amplified by the BM.

Phase effects at 50-Hz F0 were equally strong with flat-spectrum as with speech-shaped maskers. So there was no evidence that the high-frequency partials of the SP flat-spectrum masker, which were not masking any speech energy, had been recruited by a mechanism akin to CMR to facilitate the detection or the recognition of speech cues. A recent study by Buss and Grose (2009) examined the amount of CMR involved with complex signals. They showed that the CMR was large (as much as 6 dB) when the masker-target interactions resulted in compound output envelopes that were different across channels, for instance by randomizing the target partials' phase. However, the CMR was absent when the masker-

target interactions resulted in similar envelopes across channels, for instance by fixing the target partials' phase. In fact, in this case, $d\square$ is no longer cumulative across channels because the target-plus-masker samples are no longer independent, so the CMR may even be negative (−2 dB in their study). In the human voice, partials are roughly in phase, at least locally on the BM, so their interactions with the partials of a harmonic complex masker may not result in sufficiently different compound envelopes across channels. Therefore, it may not be surprising that no CMR occurs in the case of speech masked by harmonic complexes.

4.2. 100-Hz F0

At 100-Hz F0, SRTs converged for SP and RP maskers, whether the talker was male or female and whether the maskers were flat-spectrum or speech-shaped. This lack of phase effect radically contrasts with the results by Summers and Leek (1998), who used maskers at this F0. The most likely explanation for this discrepancy concerns masker level. In Summers and Leek's method, the target voice was fixed at 60, 70 or 80 dB SPL and the level of the Schroeder complexes decreased (as it started even higher) to roughly, 70, 80 and 90 dB SPL at threshold. These are intense maskers, which consisted of partial numbers 2 to 50, so the level per partial at the lowest masker level was 53 dB SPL at threshold. In comparison, the level per partial of the flat-spectrum maskers used here was 42 dB SPL for a 100-Hz F0 and the level per partial of the speech-shaped maskers ranged from 35 to 55 dB SPL. Thus the maskers used by Summers and Leek had partials whose intensity covered more of the compressive range of the input-output function. It is nonetheless surprising that this difference would result in the presence or absence of phase effects at 100-Hz F0. Another possible explanation is that SP/RP complexes are a very different pair of stimuli from positive/negative Schroeder-phase complexes, perhaps because differences in BM responses are not as large.

4.3. 200-Hz F0

Unexpectedly, at 200-Hz F0, SRTs were *lower* for the RP than for the SP masker, a result opposite to the phase effects' literature described in the introduction. However, the results are consistent with those of Gockel *et al.* (2002). They measured detection of a noise signal masked by harmonic complexes with partials in RP or in cosine phase (CP). The complexes and the noise were band-pass filtered into the same spectral region, from the tenth partial up to 5 kHz. They observed that masked detection threshold (MDT) was much lower for the CP masker than for the RP masker when the masker F0 was 62.5 Hz and this masking difference increased as masker level increased from 40 to 70 dB SPL. That pattern is consistent with the phase effects described earlier, but they also observed a much smaller effect (about 2 dB), in which MDT was *lower* for the RP masker than for the CP masker when the masker F0 was 250 Hz. This masking difference was a little more pronounced at 50 and 60 dB SPL than at 40 and 70 dB SPL. The account they offer for this small effect was related to the ability to detect changes in fluctuations of the envelope level at the output of a sliding temporal integrator (Oxenham and Moore, 1994). This integrator reflects the limited temporal resolution of the auditory system which acts as a low-pass filter. It basically smoothes the fast envelope fluctuations of the within-channel outputs shown in Figure 1. For the RP complex, the output of the sliding temporal integrator is very flat, but for the CP complex, the output of the integrator displays fluctuations at the rate of F0. At high F0s, such as 250 Hz, these fluctuations are certainly not as deep as they are at low F0s, but still apparent. Listeners may detect a noise signal by the amount of fluctuations it produces from a given background (i.e. for the masker alone): the flatter the background fluctuations, the easier the detection. For speech stimuli on the other hand, this modulation masking account does not hold so well because the masker modulation at 200 Hz would hardly mask the very slow modulations of the speech broadband envelope, typically below 10 Hz. Thus, an alternative account is needed in which the masking of high-rate modulations

of speech would be detrimental. One possibility is that these envelope periodicity cues serve for F0-segregation purposes between the naturally intonated target voice and the harmonic complex masker fixed at 200-Hz F0. The envelope periodicity cues of the target voice that provided some benefits to F0-segregation would have been more effectively masked by the pronounced modulations of the SP complex than by the flat envelopes of the RP complex.

4.4. Spectral glimpsing

Another interesting aspect of the data is that SRTs progressively decreased as the masker's F0 increased for the RP masker, but not for the SP masker. The reduction in spectral density of the maskers with doubling F0 was compensated by an increase in the level per partial, since the overall masker level remained fixed at 65 dB SPL. Despite such compensation, SRTs consistently decreased by about 3 dB for each doubling of the F0 of the RP masker. The most likely interpretation is that listeners can glimpse some target energy in spectral regions located in between masker partials. This spectral glimpsing would be facilitated as the spacing between masker partials increased and is independent of the masker phase structures. So this power-spectrum based interpretation should hold for both RP and SP maskers, but in the case of the SP masker, such a pattern was not observed, possibly because the increase in masking at lower F0s was counteracted by the large advantage of dip-listening occuring with a long fundamental period.

4.5. Ecologically relevant masking situations

The present study was motivated by a desire to examine the role of dip-listening and BM compression in more ecologically relevant situations of conversation. It is important to remember that there are many silent dips in an interfering voice occurring between syllables and between words and there are also temporal dips, that do not correspond to silences in the broadband temporal envelope, occurring in specific frequency channels at times other than during the steady-state portions of vowels. In all these temporal dips, listeners may glimpse information about the voice they aim to understand. BM compression may still be involved in such "listening in the dips" but how much it is remains unclear. The present study did not investigate these relatively long silent dips, but the extremely short dips present in steady-state harmonic maskers at the output of relatively broad filters (as shown in the bottom left panels of Figure 1). The masker phases, levels, and spectral profiles used here, were more representative of everyday listening environments than those of Summers and Leek (1998) who used flat-spectrum Schroeder-phase complexes with high masker levels. The results suggest that the benefit that can be extracted from listening in the dips of a steady-state harmonic masker may be very modest for the range of F0s produced by human speech when sound levels are moderate. This conclusion is in line with the observations of Summerfield and Assmann (1991) on simultaneous vowel sounds. Due to the periodic release of pressure at the glottis, vowel sounds are naturally well modulated at the fundamental period. When two vowels have asynchonous glottal pulses, their strong modulation at the fundamental period might be used to identify each between the glottal pulses of the other. However, they found that such pitch-period asynchrony allowed improved recognition of simultaneous synthetic vowels only at 50-Hz F0 and not at 100-Hz F0. Also in line with the present results, Deroche and Culling (2011) measured SRTs of a voice masked by speech-shaped RP and SP harmonic complexes based on a F0 of 110 Hz and presented at 69 dB SPL in anechoic and reverberant environments. The F0 of the target voice was manipulated (monotonized or sinusoidally modulated) to be 2 semitones above that of the masker. They found large impairments in SRT when the masker's F0 was sinusoidally modulated in reverberation but these impairments were not larger for SP than for RP maskers. Critically for the present topic, while reverberation flattened the masker envelope modulations of the SP masker, SRT did not increase as long as the masker's F0 remained monotonized. Therefore in their study, there was also no evidence that dip-listening was involved in the

masking efficiency of speech-shaped harmonic complexes at 110-Hz F0 presented at 69 dB SPL.

In comparison with tone detection data, it is unclear why such a rate limitation is found in the case of speech targets. There are limitations to the temporal resolution with which the BM can act, but these limitations arise at F0s above 100 Hz. For instance, Kohlrausch and Sander (1995) showed large phase effects in the MDT of a 1100-Hz pure tone for masker F0s between 50 and 200 Hz. At high F0s, such as 200 Hz, it is less clear whether listening in the dips is involved; a difference in MDT could occur simply because a peaky waveform has a reduced RMS amplitude after compression than a waveform with a flat envelope (and the same initial RMS). This explains why phase effects are also observed in forward masking conditions where listeners cannot listen in the dips (Carlyon and Datta, 1997a; Wojtczak and Oxenham, 2009). Nonetheless, BM compression is generally thought to be fast-acting (Ruggero et al., 1997), so the problem with fundamental periods as short as 10 or 5 ms is not that compression is not sufficiently fast-acting, but may rather be that too much of the speech spectrum falls in the region of spectrally resolved masker partials, for which compression cannot play much role. Shackleton and Carlyon (1994) proposed that partials be considered unresolved when the filter passes more than 3.25 partials within its 10-dB-down bandwidth. By this definition, using rounded-exponential auditory filters, partials of a 50-Hz F0 complex should be unresolved beyond 609 Hz. Thus, most of the speech spectrum, which lies beyond this frequency, is masked by unresolved masker partials at 50-Hz F0 and consequently large phase effects were observed. Partials of a 100- and 200-Hz F0 complex should be unresolved beyond 1447 Hz and 3122 Hz respectively. Although there are presumably some important cues left over in these unresolved regions, these cues may be redundant and speech may well be recognized from the region of resolved partials only, especially since listeners seem able to glimpse in between masker partials (see section D).

To sum up, there may be two main reasons for the small contribution of dip-listening for harmonic complex maskers, such as vowel sounds, at moderate sound levels. First, masking voices commonly have F0s of 100 to 200 Hz, which pushes the role of dip-listening between glottal pulses to high spectral regions (containing unresolved masker partials). In these high spectral regions, speech is less intense than it is in low spectral regions. Due to redundancy in speech, high-frequency cues might not necessarily be needed. Second, a release from masking is obviously most beneficial when there is a lot of masking to begin with. As masking voices have a great deal of energy below 1.5 kHz, mechanisms involving spectral glimpsing or the masker's periodicity (Deroche and Culling, 2011) may provide more masking release than dip-listening.

5. Summary

The present study investigated recognition of a voice masked by SP or RP harmonic complexes based on F0s of 50, 100, and 200 Hz. A substantial masking release was observed at 50-Hz F0, which was likely due to dip-listening possibly facilitated by the non-linear amplification of the basilar membrane. This masking release however did not occur at 100-Hz F0, typical of a male interfering voice. At 200-Hz F0, typical of a female interfering voice, thresholds were unexpectedly higher for SP than for RP maskers, perhaps because mechanisms related to modulation masking facilitate speech recognition in maskers with flat temporal envelopes. In conclusion, listeners do not appear to benefit from temporal dips in the within-channel envelopes of harmonic maskers with F0s in the human voice range and at a moderate sound level. Phase effects were not observed at 100- and 200-Hz F0 presumably because a large part of the speech spectrum falls in spectral regions where masker partials are generally resolved.

Acknowledgments

This work was supported by NIH Grants No. R01DC004786, No. R01DC004786-08S1, and No. R21DC011905 to M.C.

Appendix

In order to maximize the waveform differences with SP complexes, we attempted to select RP complexes that displayed flat within-channel temporal envelopes across a large range of center frequencies (CFs). We proceeded in three phases.

In phase 1, 1000 different versions of RP complexes (flat-spectrum) were generated at each F0. For each version, the complex was passed through a simulation of 256 auditory filters, regularly spaced on a ERB-scale along the entire spectrum, which consisted of rounded-exponential filters with level dependency based on the data of Glasberg and Moore (1990), and realistic phase responses based on the data of Oxenham and Dau (2001b). To limit computational time, the stimuli were restricted to 20-ms, excluding onset and offset ramps, which covered 1, 2, and 4 complete periods at each F0 respectively. For each of the 256 outputs of the filter bank, the crest factor was calculated (maximum amplitude divided by RMS amplitude). The crest factors in each channel were then averaged across the 1000 different versions to provide a mean set of 256 crest factors, i.e., a reference that was representative of the envelope modulations produced by common RP complexes across the entire spectrum. Figure A1 presents these reference crest factors (dashed lines) for each F0 as well as the crest factors of the SP complexes (solid lines) for comparison. At low CFs, the crest factors show some small variations, having local maxima when CF is in between partials. At high CFs, the crest factors show some substantial variations due to the position of CFs relative to the set of partials passing through a given filter as well as the influence of the phase curvature at this CF. More interestingly, at 50-Hz F0, the crest factors diverge between SP and RP complexes above only a few hundred Hz. At 100-Hz F0, the crest factors diverge above 1.5 kHz and at 200-Hz F0, the crest factors diverge above 3 kHz. Expectedly, this roughly corresponds to the unresolvability cut-off (Shackleton and Carlyon, 1994).

In phase 2, a new version of RP complex was generated at a given F0 and passed through the same auditory filter-bank. The crest factor was extracted for each of the 256 outputs and a t test was performed with alpha at the 0.001 significance level to determine whether this set of 256 crest factors was particularly low compared with the reference (obtained in phase 1 at the same F0). The point, here, was to choose complexes at the lower extremity of the distribution, not to test the significance of the difference, so this procedure disregarded the inflation of Type I error and was repeated at the same alpha until one complex produced significantly flatter envelopes than the reference, by chance. It took respectively, 12, 438, and 855 trials for a F0 of 50, 100, and 200 Hz to find these complexes which were chosen for the experiment. At low F0s, there can be many unresolved partials within a filter centered in high-frequency regions: different configurations in phase relationships of these unresolved partials can result in many different envelope shapes. So there is a large variability in crest factors among different RP complexes and it is easy to find a particular set of phases that lead to flat envelopes. In contrast, at high F0s, the crest factor reference (obtained in phase 1) is very representative of the population of RP complexes; it is therefore more difficult to find a RP complex that is significantly flatter than the rest of the population. That is why it took more trials to find a satisfactory RP complex at 200-Hz F0 than at 50-Hz F0. Figure A2 shows the crest factors of the RP complex at 100-Hz F0, chosen as a result of phase 2, in comparison with the reference obtained in phase 1, as well as the average of flattened complexes obtained in phase 3.

Phase 3 was concerned with engineering RP complexes with temporal envelopes as flat as possible. A gradient search procedure is efficient in finding local minima, i.e. waveforms with relatively flat envelopes, but the number of local minima grows rapidly with the number of partials (Hartmann and Pumplin, 1991). Finding the global minimum requires to restart the procedure at different points in the phase space. With as many as 440 partials, however, there are so many local minima that in practice, it is not possible to be confident about finding the global minimum. Thus, a simpler attempt was made to jitter specific phases at CFs where temporal envelopes were found to be particularly peaky. In step 1, one RP complex was generated and passed through the auditory filter bank from which 256 crest factors were extracted. In step 2, this set of crest factors was divided by the reference (obtained in phase 1) to provide normalized crest factors. This enabled the search for global maxima to not focus always on very high CFs where crest factors were generally highest. In step 3, a global maximum of these normalized crest factors was detected at a given CF. Since compression operates on the region of unresolved partials, the search for global maximum was restricted to a range of CFs above the resolvability cut-off. We then searched for the nearest partial to the CF at which the maximum was detected and changed its phase by a random value. Eight random values were tested, sampled at eight equal intervals between −pi and pi (just to ensure that no part of the phase range was completely ignored). Eight RP complex candidates were generated with the same set of phases except for the one partial that was assigned a new random phase. Each candidate was passed through the filter-bank and crest factors were extracted. The candidate that led to the largest reduction in the global maximum was selected. Step 3 was repeated recursively as a new global maximum of crest factor was detected. Figure A3 illustrates the first three reductions in global maxima resulting from this algorithm. Critically, some changes in phase can recreate global maxima at some adjacent CFs that were absent before. In other words, trying to flatten the temporal envelope at a given CF can make the temporal envelope peakier in an adjacent CF (and possibly peakier than the maximum that is being reduced). Thus, the flattening of the envelopes was a slow process that often took a few steps backwards before reaching a set of phases leading to a more homogeneous set of relatively low crest factors. When none of the eight candidates could reduce the global maximum, the procedure ended and the final set of phases provided one example of “flattened RP complex”, as illustrated on the bottom right panel of Figure A3. Phase 3 was repeated 100 times, starting with different random waveforms. The average set of crest factors for these “flattened” complexes (black dashed line) is shown in Figure A2. T tests were performed between this average set of crest factors (flattened) and the reference (obtained in phase 1): they were significant at all three F0s but the effect size was small, on average a reduction of about 0.2 in crest factor, over the range of CFs located above the resolvability cut-off. As mentioned above, the reason for this minimal reduction is that after the main maxima have been reduced, there is not much room left for further flattening. Flattening the envelope at a given CF often results in producing another peak in adjacent filters. T tests were also performed between the flattened complexes obtained in phase 3 and the complexes selected in phase 2: none was significant. In conclusion, it is possible to engineer RP complexes to have particularly flat temporal envelopes, but the reduction in envelope modulations is so small that such an algorithm is not more effective than simply generating many tokens of RP complexes until one is found to be particularly flat (as it was performed in phase 2).

List of abbreviations

BM	basilar membrane
SRT	speech reception threshold
SP	sine phase

RP	random phase
CP	cosine phase
F0	fundamental frequency
TMR	target-to-masker ratio
ms	milliseconds
SPL	sound pressure level
FIR	finite impulse response
CMR	comodulation masking release
RMS	root mean square
MDT	masked detection threshold
ERB	equivalent rectangular bandwidth
CF	center frequency

References

Alcántara JI, Moore BCJ, Glasberg BR, Wilkinson AJK, Jorasz U. Phase effects in masking: within- versus across-channel processes. J. Acoust. Soc. Am. 2003; 114:2158–2166. [PubMed: 14587613]

de Boer E, Nuttall AL. The mechanical waveform of the basilar membrane. I. Frequency modulations (‘glides’) in impulse responses and cross-correlation functions. J. Acoust. Soc. Am. 1997; 101:3583–3592. [PubMed: 9193046]

Buss E, Grose JH. Spectral integration under conditions of comodulation masking release. J. Acoust. Soc. Am. 2009; 125:1612–1621. [PubMed: 19275319]

Carlyon RP, Datta AJ. Excitation produced by Schroeder-phase complexes: Evidence for fast-acting compression in the auditory system. J. Acoust. Soc. Am. 1997a; 101:3636–3647. [PubMed: 9193051]

Carlyon RP, Datta AJ. Masking period patterns of Schroeder-phase complexes: Effects of level, number of components, and phase of flanking components. J. Acoust. Soc. Am. 1997b; 101:3648–3657. [PubMed: 9193052]

Carney LH, McDuffy MJ, Shekhter I. Frequency glides in the impulse response of low-frequency auditory-nerve fibers. J. Acoust. Soc. Am. 1999; 105:2384–2391. [PubMed: 10212419]

Deroche MLD, Culling JF. Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation. J. Acoust. Soc. Am. 2011; 130:2855–2865. [PubMed: 22087914]

Festen JM, Plomp R. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J. Acoust. Soc. Am. 1990; 88:1725–1736. [PubMed: 2262629]

Fletcher H, Galt RH. The perception of speech and its relation to telephony. J. Acoust. Soc. Am. 1950; 22:89–151.

Glasberg BR, Moore BCJ. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 1990; 47:103–138. [PubMed: 2228789]

Gockel H, Moore BCJ, Patterson RD. Asymmetry of masking between complex tones and noise: The role of temporal structure and peripheral compression. J. Acoust.Soc. Am. 2002; 111:2759–2770. [PubMed: 12083211]

Hartmann WM, Pumplin J. Periodic signals with minimal power fluctuations. J. Acoust. Soc. Am. 1991; 90:1986–1999. [PubMed: 1960291]

Hawley M, Litovsky R, Culling J. The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. J. Acoust. Soc. Am. 2004; 115:833–843. [PubMed: 15000195]

IEEE. IEEE recommended practise for speech quality measurements. IEEE Trans. Audio Electroacoust. 1969; 17:227–246.

Johnstone BM, Patuzzi RP, Yates GK. Basilar membrane measurements and the travelling wave. Hear. Res. 1986; 22:147–153. [PubMed: 3733536]

Kohlrausch A, Sander A. Phase effects in masking related to dispersion in the inner ear. II. Masking period patterns of short targets. J. Acoust. Soc. Am. 1995; 97:1817–1829. [PubMed: 7699163]

deLaat, JAPM.; Plomp, R. The reception threshold of interrupted speech. Kinke, R.; Hartman, R., editors. Physiological Bases and Psychophysics (Springer, Berlin); Hearing: 1983. p. 359-363.

Lentz JJ, Leek MR. Psychophysical estimates of cochlear phase response: masking by harmonic complexes. J. Assoc. Res. Otolaryngol. 2001; 2:408–422. [PubMed: 11833613]

Oxenham AJ, Dau T. Reconciling frequency selectivity and phase effects in masking. J. Acoust. Soc. Am. 2001a; 110:1525–1538. [PubMed: 11572363]

Oxenham AJ, Dau T. Towards a measure of auditory-filter phase response. J. Acoust.Soc. Am. 2001b; 110:3169–3178. [PubMed: 11785818]

Oxenham AJ, Dau T. Masker phase effects in normal-hearing and hearing-impaired listeners: Evidence for peripheral compression at low signal frequencies. J. Acoust.Soc. Am. 2004; 116:2248–2257. [PubMed: 15532656]

Oxenham AJ, Moore BCJ. Modeling the additivity of non-simultaneous masking. Hear. Res. 1994; 80:105–118. [PubMed: 7852196]

Plomp R, Mimpen AM. Speech-reception threshold for sentences as a function of age and noise level. J. Acoust. Soc. Am. 1979; 66:1333–1342. [PubMed: 500971]

Recio A, Rich NC, Narayan SS, Ruggero MA. Basilar membrane responses to clicks at the base of the chinchilla cochlea. J. Acoust. Soc. Am. 1998; 103:1972–1989. [PubMed: 9566320]

Ruggero MA, Rich NC, Recio A, Narayan SS, Robles L. Basilar-membrane responses to tones at the base of the chinchillacochlea. J. Acoust.Soc. Am. 1997; 101:2151–2163. [PubMed: 9104018]

Schroeder MR. Synthesis of low peak factor signals and binary sequences with low autocorrelation. IEEE Trans. Inf. Theory. 1970; 16:85–89.

Shackleton TM, Carlyon RP. The role of resolved and unresolved harmonics in pitch perception and frequency modulation. J. Acoust.Soc. Am. 1994; 95:3529–3540. [PubMed: 8046144]

Shen Y, Lentz JJ. Level dependence in behavioral measurements of auditory-filter phase characteristics. J. Acoust.Soc. Am. 2009; 126:2501–2510. [PubMed: 19894830]

Smith BK, Sieben UK, Kohlrausch A, Schroeder MR. Phase effects in masking related to dispersion in the inner ear. J. Acoust. Soc. Am. 1986; 80:1631–1637. [PubMed: 3794068]

Summerfield Q, Assmann PF. Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony. J. Acoust. Soc. Am. 1991; 89:1364–1377. [PubMed: 2030224]

Summers V, Leek MR. Masking of tones and speech by Schroeder-phase harmonic complexes in normally hearing and hearing-impaired listeners. Hear. Res. 1998; 118:139–150. [PubMed: 9606069]

Wojtczak M, Oxenham AJ. On- and off-frequency forward masking by Schroeder-phase complexes. J. Assoc. Res. Otolaryngol. 2009; 10:595–607. [PubMed: 19626368]

Yates GK. Basilar membrane nonlinearity and its influence on auditory nerve rate-intensity functions. Hear. Res. 1990; 50:145–162. [PubMed: 2076968]

Research Highlights

- SRTs were measured for an intonated voice masked by harmonic complexes with different phase structures.
- SRTs were lower for sine- than for random-phase maskers at 50-Hz F0, and vice-versa at 200-Hz F0.
- Dip-listening contributes weakly to speech recognition in harmonic complexes such as vowels.

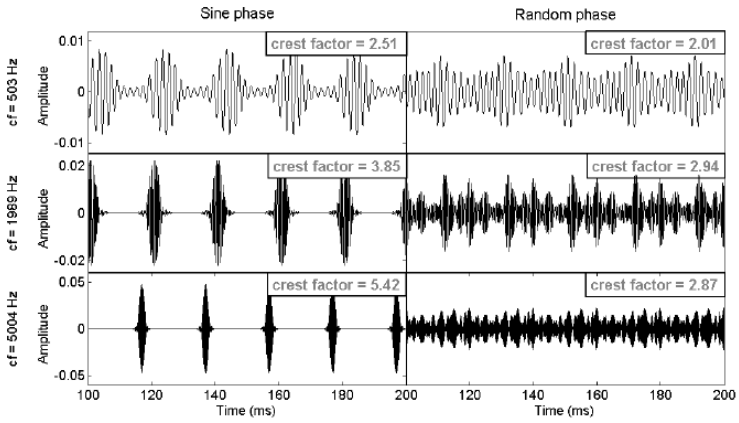

FIG 1.

Temporal waveforms of the sine-phase (left) and random-phase (right) harmonic masker based on a 50-Hz F0, at 65 dB SPL, at the output of simulated auditory filters centered at 503 (top), 1989 (middle) and 5004 (bottom) Hz.

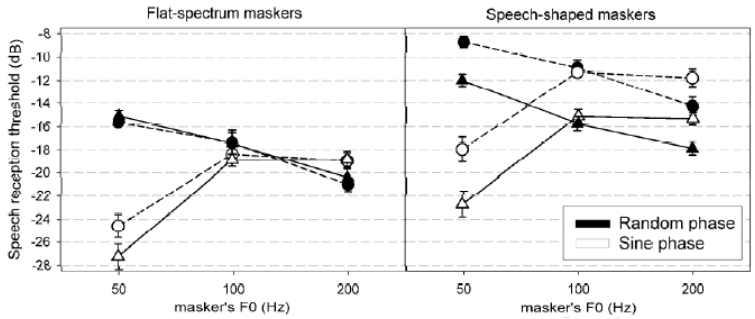

FIG 2.

Mean speech reception threshold for the flat-spectrum (left panel) and speech-shaped (right panel) SP and RP maskers as a function of the masker's F0. Target talker was male (circles) or female (triangles). Lower thresholds indicate greater intelligibility. Error bars are ±1 standard error of the mean over twelve listeners.

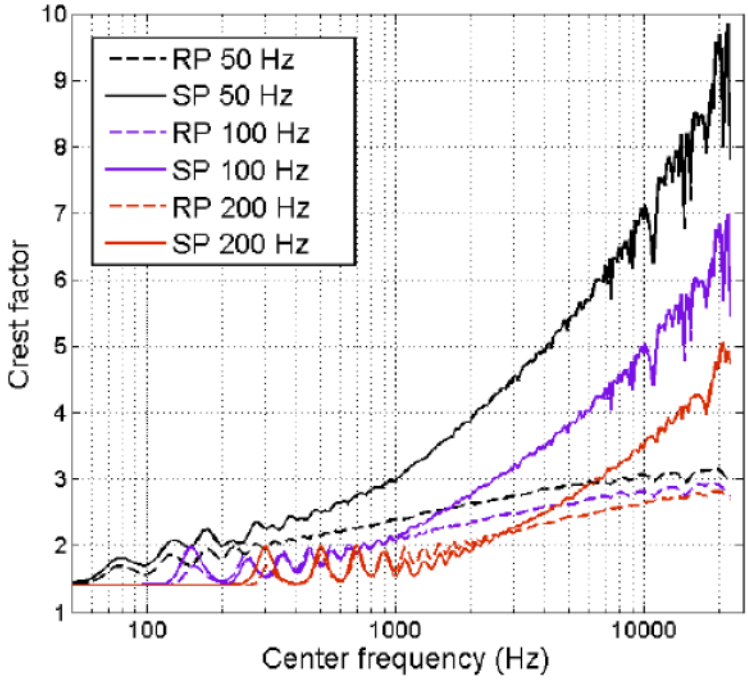

FIG A1.

Crest factors at the outputs of 256 simulated auditory filters, for the SP (solid lines) and RP (dashed lines) complexes at 50−, 100−, and 200-Hz F0. The crest factors for the RP complexes were averaged over 1000 different versions.

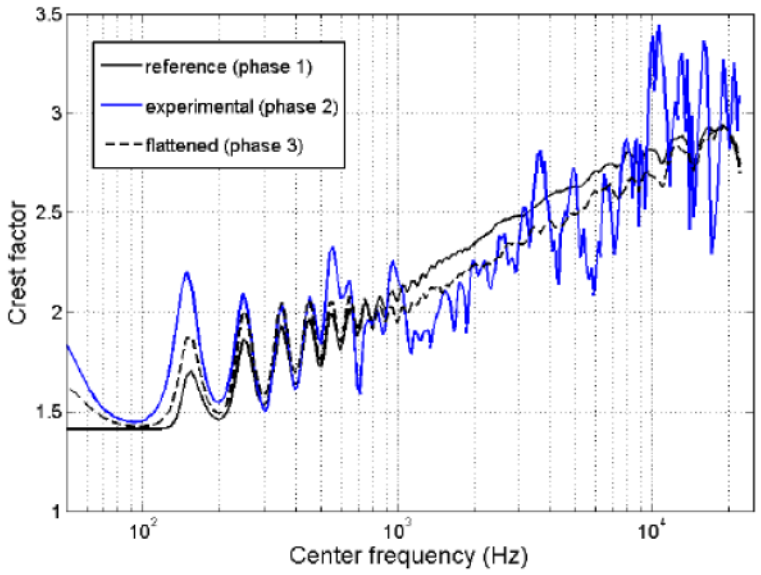

FIG A2.

Crest factors of RP complexes at 100-Hz F0 passed through the auditory filter bank. The RP complex chosen for the experiment had crest factors that were significantly lower than the reference. Flattened RP complexes were also particularly flat but not different from the RP complexes chosen in phase 2 (i.e. the experimental stimuli).

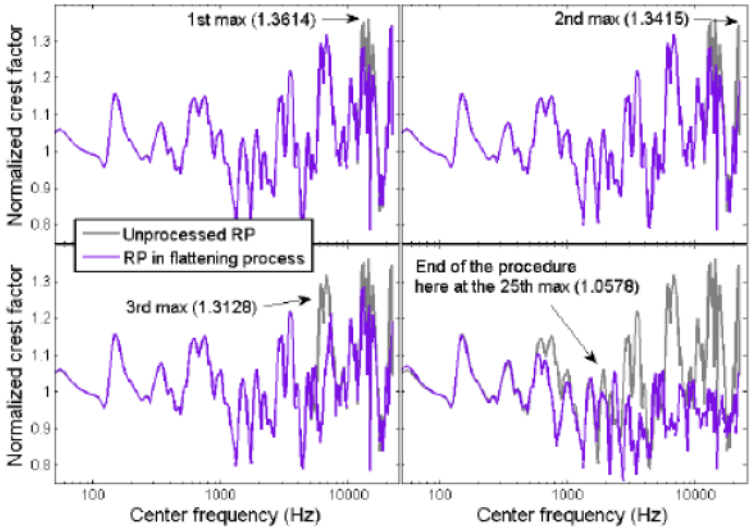

FIG A3.

Details of the flattening algorithm in the first three reductions in global maxima for a RP complex at 100-Hz F0. At the end of the procedure, a set of phases is found to provide a RP complex with particularly flat within-channel temporal envelopes in spectral regions above the resolvability cut-off.

TABLE I

Statistics of the experiment.

Factors	F value	p value
F0	F(2,44)=25.2	0.000 *
phase	F(1,22)=67.9	0.000 *
masker profile	F(1,22)=325.2	0.000 *
talker gender	F(1,22)=13.4	0.001 *
F0 × phase	F(2,44)=148.1	0.000 *
F0 × masker profile	F(2,44)=0.4	0.70
F0 × talker gender	F(2,44)=1.0	0.36
phase × masker profile	F(1,22)=9.5	0.005 *
phase × talker gender	F(1,22)=1.1	0.30
masker profile × talker gender	F(1,22)=43.2	0.000 *
F0 × phase × masker profile	F(2,44)=0.4	0.68
F0 × phase × talker gender	F(2,44)=1.7	0.20
F0 × masker profile × talker gender	F(2,44)=0.5	0.64
phase × masker profile × talker gender	F(1,22)=5.0	0.036 *
4-way	F(2,44)=0.2	0.81