Investigating the role of harmonic cancellation in speech-on-speech masking

Luna Prud'homme, Mathieu Lavandier, Virginia Best

Investigating the role of harmonic cancellation in speech-on-speech masking

Luna Prud'homme[a], Mathieu Lavandier[a,*], Virginia Best[b]

[a] *Univ. Lyon, ENTPE, Laboratoire de Tribologie et Dynamique des Systèmes UMR 5513, Rue M. Audin, 69518 Vaulx-en-Velin Cedex, France*
[b] *Department of Speech, Language and Hearing Sciences, Boston University, 635 Commonwealth Ave, Boston, MA, 02215, USA*

Abstract

This study investigated the role of harmonic cancellation in the intelligibility of speech in "cocktail party" situations. While there is evidence that harmonic cancellation plays a role in the segregation of simple harmonic sounds based on fundamental frequency (F0), its utility for mixtures of speech containing non-stationary F0s and unvoiced segments is unclear. Here we focused on the energetic masking of speech targets caused by competing speech maskers. Speech reception thresholds were measured using seven maskers: speech-shaped noise, monotonized and intonated harmonic complexes, monotonized speech, noise-vocoded speech, reversed speech and natural speech. These maskers enabled an estimate of how the masking potential of speech is influenced by harmonic structure, amplitude modulation and variations in F0 over time. Measured speech reception thresholds were compared to the predictions of two computational models, with and without a harmonic cancellation component. Overall, the results suggest a minor role of harmonic cancellation in reducing energetic masking in speech mixtures.

Keywords: Speech Intelligibility, Harmonic Cancellation, Auditory Modeling, Binaural Hearing

*Corresponding author
Email address: mathieu.lavandier@entpe.fr (Mathieu Lavandier)

1. Introduction

In "cocktail party" scenarios (Cherry, 1953), where speech must be understood in the presence of noise and competing talkers, there are several factors that may improve intelligibility, including spatial separation of sources, masker amplitude modulation, and differences in harmonic structure between sources. While the effects of spatial separation and masker amplitude modulation have been extensively characterized and incorporated into computational models of speech intelligibility (Lavandier & Best, 2020), the role of harmonicity is less well understood.

In the context of speech-on-speech masking, several previous studies showed a beneficial effect of fundamental frequency (F0) differences. Brokx & Nooteboom (1982) tested the intelligibility of monotonized target speech against monotonized masker speech. They found that the percentage of errors decreased with increasing F0 difference between target and masker, except when the difference was one octave. Darwin et al. (2003) found that differences in F0 and vocal tract length both improved segregation of target and masker talkers. In a recent study, Popham et al. (2018) tested speech intelligibility with target and masker that were either both natural (harmonic) speech or both inharmonic speech. The percentage of correctly reported words decreased when the speech was inharmonic, suggesting that harmonicity is important for the successful segregation of speech mixtures. In each of these studies, it is not entirely clear what mechanisms underlie the observed benefits. One of the complicating factors is that speech-on-speech situations involve both energetic masking (EM, Culling, 2016) and informational masking (IM, Kidd et al., 2016). Whereas EM refers to masking that renders target speech inaudible, IM refers to masking that happens even when the target speech is audible, and is often attributed to the listener's inability to perceptually segregate competing voices and selectively attend to the target talker. Harmonicity-based effects could theoretically reduce both kinds of masking. For example, harmonic structures may be readily suppressed by the auditory system, thus reducing EM. Alternatively, differences in

F0 may enhance the perceptual segregation of competing talkers and effectively reduce IM.

The present study focuses on harmonicity-based effects on EM, and deliberately puts aside harmonicity-based effects on IM. A handful of recent studies are highly relevant here as they measured speech intelligibility in the presence of harmonic complexes (often called buzzes; Leclère et al., 2017; Steinmetzger & Rosen, 2015; Deroche & Culling, 2013). These stimuli have the benefit that they are harmonic, but they are not speech-like and thus they can be assumed to cause primarily EM. These studies demonstrated that harmonic sounds exert less EM than inharmonic sounds, and considered several possible mechanisms. Deroche & Culling (2013) suggested that listeners might be able to extract target information in the spectral dips of harmonic maskers, between the resolved partials ("spectral glimpsing"). Another possibility is that listeners may detect the harmonic structure of the masker and suppress it when its F0 is different from the one of the target (harmonic cancellation, de Cheveigné, 2021, 1993). It is unclear though whether this reduction in EM translates to natural sounds like speech. With speech maskers, there are several co-existing acoustic characteristics to consider that may complicate the picture: the presence of unvoiced segments, intonation, and amplitude modulation. Indeed, there are some indications that these factors may reduce the benefits associated with harmonic maskers (Deroche & Culling, 2013; Leclère et al., 2017; Deroche & Gracco, 2019). For example, Leclère et al. (2017) showed that the advantage due to harmonicity was reduced for intonated buzzes compared to monotonized buzzes. Deroche & Gracco (2019) tested speech intelligibility against harmonic complexes and monotonized speech maskers with one or two F0s. They found different results for the two types of masker, suggesting that F0-based unmasking might operate differently for speech and buzz maskers.

The current study aimed to extend these findings by comparing speech intelligibility for a range of harmonic and inharmonic maskers (including noise, buzzes and speech) in the same group of listeners under otherwise identical conditions. By comparing performance across masker types, the aim was to

estimate the influence of harmonicity while controlling for various other characteristics of speech. IM was deliberately minimized in order to focus on energetic aspects of harmonicity-based unmasking.

The present study is also part of a larger effort to develop a speech intelligibility model that accurately captures the EM caused by complex interfering sounds such as speech. Our motivation for doing so is to have a means by which the relative contributions of EM and IM can be confidently estimated in real-world listening situations. Previous studies have estimated EM caused by speech maskers using models validated only for amplitude-modulated noise maskers, thus implicitly (Biberger & Ewert, 2019) or explicitly (Lavandier et al., 2021; Wasiuk et al., 2020) assuming that masker harmonicity was not relevant. Steinmetzger et al. (2019) tested four models on speech intelligibility data collected in the presence of speech-shaped noise and intonated harmonic complexes (Steinmetzger & Rosen, 2015). All of the models underestimated the benefit due to masker harmonicity. The authors speculated that the performance of the models might be improved by implementing a mechanism of enhanced stream segregation dependent on masker harmonicity. Prud'homme et al. (2020) proposed a model with an implementation of harmonic cancellation that allowed good predictions of intelligibility against stationary buzzes. This model was further developed to take into account the detrimental effect of masker intonation by Prud'homme et al. (2022), but was never tested in speech-on-speech scenarios. In the present study, by combining behavioral measurements and model predictions, we aimed to investigate whether harmonicity (and harmonic cancellation in particular) plays a role in reducing EM in speech-on-speech situations. Specifically, we compared the predictions of models with and without harmonic cancellation, to determine whether this mechanism is a necessary component of a comprehensive speech intelligibility model.

2. Main experiment

2.1. Methods

2.1.1. Listeners

Nine listeners (ages 19-23 years, mean age 21) participated in the main experiment (ten listeners originally participated but one listener had unusually poor speech intelligibility, even in quiet, and thus their data were excluded from the analysis). All listeners had normal pure tone thresholds (not exceeding 20 dB HL at octave frequencies from 250 to 8000 Hz) and were paid for their participation. All procedures were approved by the Boston University Institutional Review Board.

2.1.2. Stimuli

The target sentences were matrix sentences taken from a closed set (Kidd et al., 2008). The sentences consisted of five words (name, verb, number, adjective, object) and each word was drawn randomly from a set of eight options. The target sentence was always spoken by the same North American female voice (mean F0 = 180 Hz).

Seven types of masker were used: speech-shaped noise (SSN), monotonized buzz, intonated buzz, monotonized speech, natural speech, vocoded speech and reversed speech. SSN was used as a baseline without any harmonicity or amplitude modulation. The monotonized and intonated buzzes are non-speech maskers without the slow amplitude modulations that are characteristic of speech, but with harmonicity (F0 fixed or varying over time, respectively). Vocoded speech has amplitude modulations that are similar to speech, which should provide a similar amount of temporal dip listening, but no harmonicity. Monotonized speech was included because previous data suggested that harmonic cancellation operates more effectively for monotonized than intonated buzz maskers (Leclère et al., 2017). Reversed speech was added as a control to confirm that our forward (natural) speech masker caused minimal IM: since reversed speech causes very little IM then we expected no difference between these two maskers.

The seven maskers were all derived from the same speech monologue spoken by an Australian accented male talker (mean F0 = 112 Hz). The SSN was a white noise filtered to have the same long-term excitation pattern (Glasberg & Moore, 1990) as the monologue. The buzzes were harmonic complexes with partials in random phase. The buzz was either monotonized with a fixed F0 of 112 Hz, or intonated with a continuous F0 contour extracted from the speech monologue. This F0 contour was applied to the buzz using PRAAT PSOLA (Boersma & Weenink, 2018). The monotonized speech was created by fixing the F0 of the monologue to its mean (112 Hz) using PRAAT PSOLA. The vocoded speech was created with an 8-channel vocoder, using an envelope low-pass filter-cutoff frequency of 150 Hz. All maskers were passed through a 0.1-s finite impulse response filter to match the average long-term excitation pattern of the monologue. Figure 1 presents the excitation patterns of all maskers.

Stimuli were spatialized using anechoic KEMAR head related transfer-functions (HRTFs; Gardner & Martin, 1995). The target was presented at 0° azimuth and the masker was either presented at 0° azimuth (co-located condition) or 60° to the side (separated condition). This difference in location, as well as the difference in talker sex, and the difference in the structure and content of the speech materials, all served to distinguish the target from the masker and thus minimize IM.

2.1.3. Procedure

The stimuli were presented via a 24-bit soundcard (RME HDSP 9632, Haimhausen, Germany) to a pair of circumaural headphones (Sennheiser HD 280 Pro, Wedemark, Germany). The listeners were seated in a double-walled sound treated booth. After listening to each masked sentence, they were presented with a grid of 40 words (five categories x eight options). They were instructed to select one word in each category and were then presented with correct answer feedback.

The experiment took two sessions of approximately two hours each. Each session was composed of twenty-nine blocks of 20 trials. Within a block, the masking condition was fixed but the signal-to-noise ratio (SNR) varied randomly

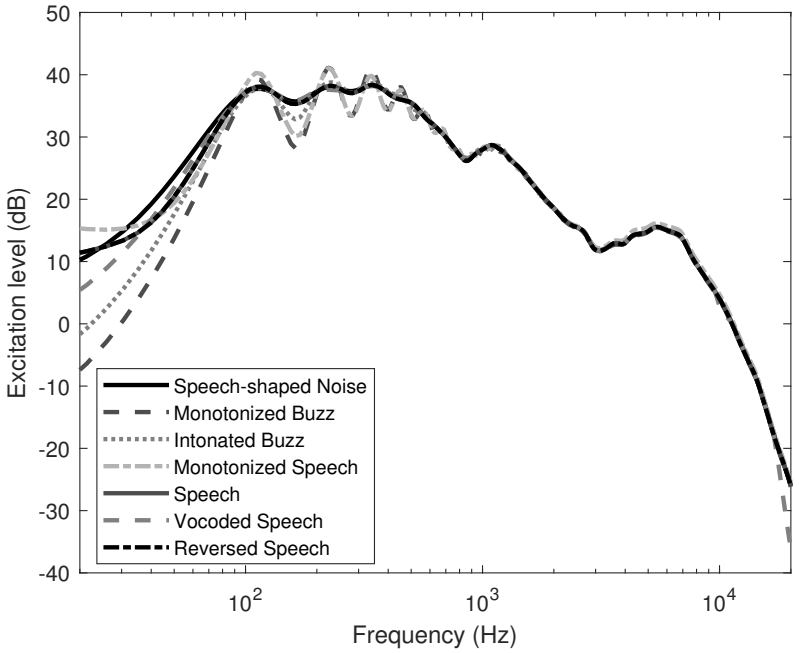

Figure 1: Long-term excitation patterns of the seven maskers tested in the main experiment

between five chosen values (from -40 to -10 dB). The masker presentation level was fixed at 65 dB SPL and the target level was varied to achieve the required SNR. The first block in each session consisted of target sentences alone (in quiet) to familiarize the listeners with the task and to make sure that they were able to understand the speech in quiet at the different target levels (which ranged from 25 to 55 dB SPL). This was followed by twenty-eight blocks of testing. Each masking condition was presented twice per session. The order was randomized across participants. At the end of the two sessions, each listener had performed four blocks in each condition, which resulted in 80 scored words at each SNR.

The percentage of correct words was calculated for each participant at each SNR in each condition. Logistic functions were fitted using the psignfit toolbox version 4 for MATLAB, which implements the maximum likelihood method described by Wichmann & Hill (2001). The lower asymptote was set to chance performance (12.5%). Speech reception thresholds (SRTs) corresponding to 50% correct were extracted from the logistic fits.

2.2. Results

All subjects (except the one whose data was excluded from the analysis) performed well in quiet with scores above 85% correct at all target levels.

Figure 2, top panel, presents the mean SRTs across listeners for the different masker types in the co-located and separated conditions. A repeated-measures ANOVA was performed with two factors (masker type x spatial separation). The main effects of spatial separation, masker type and their interaction were significant ($F(1,8) = 940.99$, $p<0.001$, $F(6,48) = 107.91$, $p<0.001$ and $F(6,48) = 15.56$, $p<0.001$, respectively). Post hoc paired t-tests ($p<0.05$) suggested that in the co-located condition, SRTs were significantly higher for SSN than for the buzzes, and were significantly higher for intonated than monotonized buzzes. A similar pattern was found in the separated condition, with the exception that the difference in SRT between SSN and intonated buzz was not significant. In both spatial conditions, there was no significant difference in SRT between natural speech and the other speech-based maskers (monotonized speech, vocoded

speech, reversed speech). The important results here are that (a) the effect of intonation observed for the buzz maskers was not replicated for the speech maskers, (b) the absence of periodicity in the vocoded speech masker had little effect on SRTs, and (c) forward and reversed speech maskers gave similar SRTs consistent with minimal IM.

3. Control experiments

3.1. Rationale

SRTs in the main experiment were very low (-32 dB on average in the separated conditions), presumably because of the large number of cues available to reduce both EM and IM. Because of these low values, there was some concern that a floor effect may have limited our ability to see differences across masker types. Thus, two control experiments were conducted to increase the SRTs in different ways. In control experiment 1, the speech task was made more difficult by using open-set target sentences as opposed to closed-set matrix sentences. In control experiment 2, in addition to using open-set materials, a second speech masker was added to increase the overall amount of masking.

In addition to increasing the overall difficulty, the idea behind control experiment 2 was that increasing the EM may increase the opportunities for harmonic cancellation to operate. Specifically, by adding a second speech masker, there should be more voiced parts and fewer dips overall in the speech masker. This experiment was designed so that IM, which is known to be more prominent for two-talker maskers than for single-talker maskers (Freyman et al., 2004; Brungart et al., 2001), was still minimized as much as possible.

3.2. Methods

Five listeners (ages 19-22 years, mean age 21) who participated in the main experiment also participated in control experiment 1. Five new listeners who did not participate in the main experiment (ages 19-22 years, mean age 21; normal pure tone thresholds at octave frequencies from 250 to 8000 Hz) participated in control experiment 2.

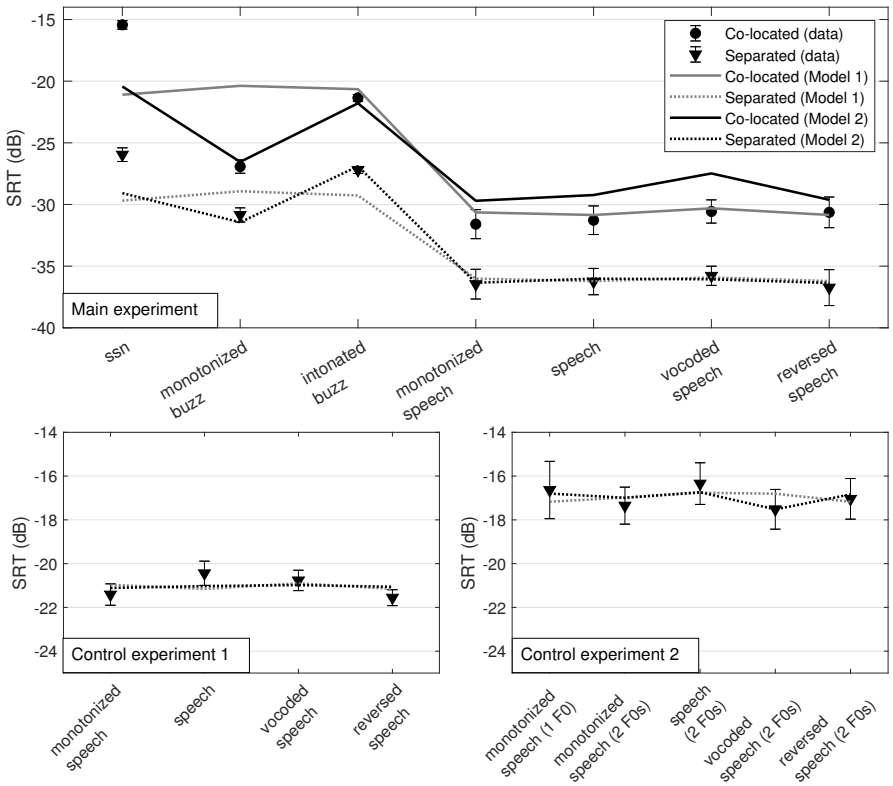

Figure 2: Mean SRTs with standard errors across participants measured in the main experiment (top panel), in the control experiment 1 (bottom left panel), and the control experiment 2 (bottom right panel) with the corresponding model predictions using: model 1, a binaural model without harmonic cancellation (Vicente & Lavandier, 2020) and model 2, a binaural model with harmonic cancellation (Prud'homme et al., 2022).

In the two control experiments, the target sentences were taken from the Harvard Sentence List (Rothauser et al., 1969) and were composed of five keywords. They were spoken by a female talker with a North American accent (mean F0 = 190 Hz). Only the spatially separated configuration was tested, and in control experiment 2, the two maskers were both presented at the same location (60° azimuth).

In control experiment 1, only four masker conditions (those that could have been influenced by a floor effect) were re-tested: speech, monotonized speech, vocoded speech and reversed speech.

In control experiment 2, those same masker conditions were tested with two maskers instead of one. The maskers were generated as in the main experiment, derived from speech monologues spoken by male voices (one at a mean F0 of 112 Hz, same monologue as the main experiment, the other at a mean F0 of 130 Hz), both with an Australian accent. This resulted in four masker types: two-talker natural speech, two-talker monotonized speech, two-talker vocoded speech, two-talker reversed speech. A fifth masker type was constructed by adding two time-shifted copies of the monotonized male monologue from the main experiment. This resulted in a two-talker masker with a single steady F0 at 112 Hz. This masker was theoretically optimized for harmonic cancellation, having a higher proportion of voiced energy than the single-talker version and a single, steady F0.

In the control experiments, the participants were instructed to type the sentence they heard. The correct transcript was then displayed on the screen, with the keywords in capital letters and the participants had to self-mark their number of correct keywords. For each control experiment, the listeners performed two sessions of one hour. Each session was composed of 20 blocks of ten sentences in one of the conditions and at one SNR. Five SNRs were tested from -30 to -10 dB. The listeners also performed a block in quiet at the start of each session. SRTs were extracted as per the main experiment.

3.3. Results

Figure 2 presents the SRTs measured in control experiments 1 and 2, in the bottom left and right panels, respectively. The SRTs in the control experiments were higher than in the main experiment as intended. Repeated measures ANOVAs found no significant effect of masker type in control experiment 1 [$F(3,12) = 1.54$, $p = 0.26$] or control experiment 2 [$F(4,16) = 1.79$, $p = 0.18$]. These results corroborate the key results of the main experiment, again suggesting that the harmonic structure of a speech masker has little impact on SRTs under conditions of minimal IM.

4. Modeling

4.1. Rationale

In order to further investigate the potential role of harmonic cancellation in the present study, two speech intelligibility models were applied to the stimuli. The chosen models were the binaural model proposed by Vicente & Lavandier (2020) and validated only for amplitude-modulated noise maskers, and the binaural harmonic cancellation model proposed by Prud'homme et al. (2022) validated only for harmonic complex maskers (with both intonation and amplitude modulation). By comparing how well each model can account for the data, we aimed to provide further support for (or against) a role of harmonic cancellation in reducing EM for speech maskers.

4.2. Models

The models tested here have the same structure. The inputs for both models are the target and masker signals at the ears of the listener. The target input is represented by an averaged signal obtained by adding several target sentences (Vicente & Lavandier, 2020). The masker signal is segmented into time frames using half-overlapping Hann windows. The signals are passed through a Gammatone filterbank (Patterson et al., 1987) and a SNR is computed in each frequency band.

In model 1, the advantages due to better-ear listening and binaural unmasking are computed in parallel. In each time frame and frequency band, the best SNR across the two ears is selected to obtain the better-ear SNR. The time frame duration used for the better-ear SNR computation is 24 ms. A 20 dB ceiling , which corresponds to the maximum SNR allowed in each frequency band and time frame, is introduced in order to prevent the SNR to tend to infinity in the temporal gaps of the masker. In each frequency band and time frame the binaural unmasking advantage is computed using an equation proposed by Culling et al. (2005) to estimate the binaural masking level difference (BMLD). This value is computed using a time frame duration of 300 ms. The binaural unmasking advantage and the better-ear SNR are then integrated across frequency using the SII weightings (ANSI S3.5, 1997) and averaged across time frames. The two values are added to obtain an effective SNR. This model is able to accurately predict spatial release from masking (SRM) and dip listening for speech presented against stationary and modulated noise maskers (Vicente & Lavandier, 2020).

Model 2 was proposed by Prud'homme et al. (2022) and incorporates harmonic cancellation. The masker is segmented into 300-ms time frames and the mean F0 over the time frame is computed using PRAAT PSOLA (Boersma & Weenink, 2018). A comb filter tuned to the masker F0 is applied to both target and masker signals in order to simulate the mechanism of harmonic cancellation. The harmonic cancellation component of the model adopted the same parameters suggested by Prud'homme et al. (2020, 2022): a jitter in the estimation of the F0 (0.25F0), the width of the notches of the comb filter (0.6F0), an SNR ceiling (40 dB), and a frequency limit up to which harmonic cancellation is applied (5000 Hz). The better-ear SNR is computed (as in model 1) for both the comb-filtered and unfiltered signals. The best better-ear SNR between the two is chosen (i.e., harmonic cancellation is only applied if it provides an advantage). Another condition for applying harmonic cancellation is that the masker signal is voiced at least 50 % of the time in the considered time frame. If this is not the case, the better-ear SNR and binaural unmasking advantage are computed

as in model 1. In model 2, harmonic cancellation and binaural unmasking are mutually exclusive (if harmonic cancellation is active, only the better-ear SNR is computed). The effective SNR is obtained by integrating the binaural unmasking advantage and/or the better-ear SNR across frequency bands using the SII weightings, and averaging them across time frames. As per Prud'homme et al. (2020, 2022), to obtain the model predictions the model is run 800 times for each condition using a different realization of the stimuli and a different value of the jitter (drawn from a normal distribution with a standard deviation corresponding to 0.25F0). The random jitter leads to a different prediction on each of these "trials", which are then averaged to obtain the final prediction. Model 2 proved useful to predict the effects of spatial separation, intonation and amplitude modulation for speech intelligibility against buzz maskers (Prud'homme et al., 2022).

Both models are able to predict relative differences between SRTs in different conditions, but cannot provide an absolute prediction of intelligibility. To do so, a reference needs to be chosen, which is typically the average SRT across conditions (Lavandier et al., 2012). The models were applied to the stimuli of the three experiments of the present study. The reference chosen to fit the predictions to the data was the average SRT across conditions for each experiment.

4.3. Results

The gray lines in Figure 2 (top panel) present the model predictions for the main experiment using model 1. This model does not predict any SRT difference between SSN, monotonized buzzes and intonated buzzes. It underestimates the difference in SRTs between SSN and the amplitude-modulated maskers (i.e., the effect of dip listening). The model accurately predicts the SRM for SSN and amplitude-modulated maskers, but it overestimates the SRM for buzz maskers. The black lines in Figure 2 (top panel) present the model predictions using model 2. This model accurately predicts the SRT differences between intonated and monotonized buzzes. It accurately predicts the SRT differences between buzzes and amplitude-modulated maskers. However, it underestimates the SRT

differences between SSN and the other masker types. The predicted SRM is similar to that observed in the data, although it is slightly overestimated for amplitude-modulated maskers.

Figure 2 (bottom panels) also presents the model predictions for the control experiments using model 1 (gray lines) and model 2 (black lines). Both models produce very similar predictions across conditions (less than 1 dB of variation), consistent with the lack of any significant differences in the behavioral data.

5. Discussion

5.1. Harmonicity

In the main experiment, SRTs were highest for the SSN masker, which had neither harmonicity nor the slow amplitude modulations that might support temporal dip listening. SRTs were 5.9 and 11.5 dB higher for SSN compared to intonated and monotonized buzzes, respectively, in the co-located condition. This result is consistent with those of Steinmetzger & Rosen (2015), who measured SRTs for speech against SSN and intonated harmonic complexes and found that SRTs were higher for SSN. The overall difference suggests that there is a harmonicity-based benefit that could be due to harmonic cancellation and/or spectral glimpsing. Another possibility is that periodic sounds cause less "modulation masking" than aperiodic sounds (Stone et al., 2011; Steinmetzger et al., 2019).

Model 1, which should capture spectral glimpsing (by computing SNR within frequency bands), predicts no differences between SSN, monotonized buzzes and intonated buzzes, suggesting that differences in spectral glimpsing are negligible in terms of SNR. Model 2, which incorporates harmonic cancellation, is able to predict the difference between monotonized and intonated buzzes, consistent with the results of Prud'homme et al. (2022). However, model 2 underestimates the difference between SSN and buzzes by 4.6 dB. The only previous study that tried to predict the SRT difference between noise and intonated buzz, using a modulation-based model, also underestimated this effect by about 5 dB (Stein-

metzger et al., 2019). It seems that a combination of harmonic cancellation and a mechanism that is sensitive to modulation masking may be required to fully predict the benefits of harmonicity for simple maskers.

The finding of higher SRTs for intonated compared to monotonized buzzes is consistent with a number of previous studies (Deroche & Culling, 2011; Green & Rosen, 2013; Leclère et al., 2017). As proposed by Leclère et al. (2017), the difference between intonated and monotonized maskers could be due to limitations in the harmonic cancellation mechanism. Specifically, it may be that it is difficult to "follow" the F0 contour when it varies over time. Another possibility is that the concomitant amplitude fluctuations that accompany intonation increase modulation masking. It is also interesting to note that in the separated conditions, the SRT difference between SSN and intonated buzz was greatly reduced, to the point that there was no significant advantage of harmonicity. It is possible that spatial separation provided enough masking release so that there was no further benefit to be gained from the weak harmonicity cue in the intonated condition.

No significant difference in SRT could be found between the speech and monotonized speech maskers in any of the experiments. Given the robust effects of intonation observed for buzzes, one might have expected a similar effect to be observed for speech (i.e., lower SRTs for a monotonized speech masker than for a naturally intonated speech masker). If the effect of intonation observed for buzzes is due to harmonic cancellation, it apparently does not apply to speech maskers. Another interesting result was that model 2 did not *predict* an advantage for monotonized speech compared to naturally intonated speech, contrary to the advantage it predicted for monotonized buzzes over intonated buzzes. There was also no significant difference between the SRTs for speech and vocoded speech. Given that the main difference between these two maskers is harmonicity, this result provides a further indication that harmonicity does not strongly affect the EM present in speech-on-speech situations. This result can be compared to that of Rosen et al. (2013), who made a similar comparison but did not attempt to reduce IM. In their one- and two-talker masker conditions,

performance was consistently poorer for speech maskers than for vocoded-speech maskers (by about 3-7 dB). The fact that we estimated no differences in EM bolsters the conclusion that their effects were related to IM.

So why might harmonic cancellation be so ineffective at reducing EM for speech maskers? Our results point to the presence of unvoiced parts and/or amplitude modulation in speech. The modeling results of Prud'homme et al. (2022) with amplitude-modulated buzz maskers suggest that harmonic cancellation can still be useful on amplitude-modulated signals. Thus, the most parsimonious explanation for the lack of an effect of harmonic cancellation with speech maskers may be the presence of unvoiced segments. Or perhaps it is a combination of these effects. For example, if the target is primarily understood based on information available in the temporal dips of a modulated masker, and for speech maskers these low energy dips tend to be unvoiced rather than voiced, this could explain why harmonic cancellation has little effect.

One concern with the absence of significant differences in SRT between the different amplitude-modulated maskers in the main experiment was that the SRTs were very low, and could have been affected by a floor effect. However, the results from the two control experiments lessen this concern: measured SRTs were substantially higher and the SRT differences between the four amplitude-modulated masker conditions were still not significant. In control experiment 2, our hypothesis was that if harmonic cancellation was at play, it would operate most effectively for the monotonized speech masker with a single F0 as it should be the easiest to cancel. The results do not confirm this hypothesis: the SRT for this monotonized speech masker was not significantly different from that of any other masker type.

Like in the main experiment, the control experiments revealed no significant differences between SRTs for vocoded speech (with no harmonic structure) and natural or monotonized speech. This suggests once again that harmonicity in the masker did not play an important role here, and the model predictions are consistent with that hypothesis. Of course, given the low number of participants in these experiments, we cannot provide definitive evidence for the lack of an

effect of masker type. However, we assume that if there is any effect, it is very small.

Our conclusion is that harmonicity-based effects on EM may be negligible for speech-on-speech situations, and that examples of harmonicity-based release from masking reported in the literature for speech maskers (Brokx & Nooteboom, 1982; Deroche & Gracco, 2019; Popham et al., 2018) likely reflect a release from IM. From a modeling perspective, this suggests that predictions of EM in speech-on-speech situations might not need to take into account the effects of harmonicity, at least as a first approximation, even if it has been shown to be important for masking caused by harmonic complexes (Prud'homme et al., 2022). In other words, for predicting the EM present in cocktail party situations, the "modulated-noise" models such as those proposed by Vicente & Lavandier (2020) or Beutelmann et al. (2010) are likely to be sufficient.

5.2. Spatial separation

As expected, we observed a main effect of spatial separation between target and masker, and the general pattern of SRM across masker types was consistent with previous studies (Beutelmann & Brand, 2006; Jelfs et al., 2011; Leclère et al., 2017; Culling & Lavandier, 2021).

The SRM predicted by model 1 is equivalent for SSN and buzzes, because this model cannot distinguish buzzes from noises apart from their minor differences in long-term spectrum (Fig. 1). This predicted SRM is similar to that observed in the data for SSN, but larger than that observed for buzzes. For amplitude-modulated maskers, the predicted SRM is in good agreement with the observed SRM. The reason that the model predicts less SRM for amplitude-modulated maskers appears to be due to the SNR ceiling set at 20 dB; as the dip listening already provides a large SNR advantage, there is limited headroom for further advantages.

Model 2 predicts the SRM reasonably well. Importantly, the success of this model relies on the fact that harmonic cancellation and binaural unmasking are mutually exclusive (Prud'homme et al., 2022). Without this assumption, the

model would overpredict SRM for harmonic maskers. The rationale behind this assumption is that binaural unmasking is a release from simultaneous masking and relies on the spectrotemporal overlap of the target and masker signals. Thus, binaural unmasking should be larger for noise than for spectrally sparse maskers like buzzes. Additional experimental work focused specifically on the interaction between binaural unmasking and harmonic cancellation would be needed for a more complete picture.

5.3. Amplitude modulation

All amplitude-modulated maskers (speech, monotonized speech, vocoded speech, reversed speech) produced lower SRTs than the other masker types. The difference in SRTs between SSN and vocoded speech was almost 15 dB in the present study, which is larger than previously reported using similar stimuli (e.g. Beutelmann et al., 2010; Collin & Lavandier, 2013). It may simply be that our vocoded speech masker contained more or larger temporal dips than the maskers used in previous studies. There is also some evidence that the benefit derived from amplitude fluctuations in the masker depends on the nature of the target speech materials (Schoof & Rosen, 2015; Best et al., 2019), and thus the discrepancy may be explained by the use of matrix sentences (which are potentially easier to guess than open-set materials) in our experiment. Indeed model 1 has been shown to accurately predict dip listening benefits measured with open-set sentences (Vicente et al., 2020), but it underpredicts the benefit observed here with matrix sentences.

Leclère et al. (2017) found that masking release due to amplitude modulation was small for buzzes, to the point that there was no advantage for monotonized buzzes with amplitude modulations compared to stationary monotonized buzzes. Steinmetzger & Rosen (2015) also found that the advantage due to amplitude modulation was smaller than the advantage due to periodicity. The present results provide a slightly different picture for speech maskers, in that evidence was found for temporal dip listening but no strong evidence for release due to harmonicity. It is possible that the relative contribution of these two

mechanisms depends on the specific masker signal characteristics. For example, harmonic cancellation may dominate for buzzes that have a rich harmonic structure, whereas dip listening may dominate for partially harmonic/voiced stimuli such as speech.

6. Conclusion

SRTs were measured for maskers ranging from noise to speech in order to better understand harmonicity-based contributions to EM in speech-on-speech situations. The different masker types provided a comparison between maskers with and without harmonic structure, amplitude modulation and variations in F0 over time.

SRTs measured for unmodulated maskers (SSN and buzzes) suggest an advantage due to harmonicity in the masker that is impaired by intonation. Such conditions continue to provide the most compelling case for a harmonic cancellation mechanism. On the other hand, the results for various amplitude-modulated "speech-like" maskers suggest a very limited role for harmonic cancellation in reducing EM in mixtures of talkers. This conclusion is further supported by the predictions of models with and without a harmonic cancellation component. Overall, we suggest that models validated for amplitude-modulated noise maskers can be used as a first approximation for predicting EM in cocktail party scenarios.

7. Acknowledgments

This work was performed within the LabEx CeLyA (Grant No. ANR-10-LABX-0060) and funded by the "Fondation Pour l'Audition" (Speech2Ears grant). V.B. was supported, in part, by National Institutes of Health-National Institute on Deafness and Other Communication Disorders (NIH-NIDCD) Award No. DC015760.

References

References

ANSI S3.5 (1997). Methods for Calculation of the Speech Intelligibility Index. *American National Standards Institute, New York*, .

Best, V., Roverud, E., Baltzell, L., Rennies, J., & Lavandier, M. (2019). The importance of a broad bandwidth for understanding "glimpsed" speech. *The Journal of the Acoustical Society of America*, *146*, 3215. doi:10.1121/1.5131651.

Beutelmann, R., & Brand, T. (2006). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *120*, 331–342. doi:10.1121/1.2202888.

Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *The Journal of the Acoustical Society of America*, *127*, 2479–2497. doi:10.1121/1.3295575.

Biberger, T., & Ewert, S. D. (2019). The effect of room acoustical parameters on speech reception thresholds and spatial release from masking. *The Journal of the Acoustical Society of America*, *146*, 2188–2200. doi:10.1121/1.5126694.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.42, retrieved 15 August 2018 from http://www.praat.org/.

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

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*, 2527–2538. doi:10.1121/1.1408946.

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*, 975–979. doi:10.1121/1.1907229.

Collin, B., & Lavandier, M. (2013). Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers. *The Journal of the Acoustical Society of America*, *134*, 1146–1159. doi:10.1121/1.4812248.

Culling, J. F. (2016). Speech intelligibility in virtual restaurants. *The Journal of the Acoustical Society of America*, *140*, 2418–2426. doi:10.1121/1.4964401.

Culling, J. F., Hawley, M. L., & Litovsky, R. Y. (2005). Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)]. *The Journal of the Acoustical Society of America*, *118*, 552–552. doi:10.1121/1.1925967.

Culling, J. F., & Lavandier, M. (2021). Binaural unmasking and spatial release from masking. In R. Y. Litovsky, M. J. Goupell, A. N. Popper, & R. R. Fay (Eds.), *Binaural Hearing* (pp. 209–241). Springer Nature Switzerland volume 73 of *Springer Handbook of Auditory Research*.

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *114*, 2913. doi:10.1121/1.1616924.

de Cheveigné, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *The Journal of the Acoustical Society of America*, *93*, 3271–3290. doi:10.1121/1.405712.

de Cheveigné, A. (2021). Harmonic Cancellation—A Fundamental of Au-

ditory Scene Analysis. *Trends in Hearing*, *25*, 2331–2165. doi:10.1177/23312165211041422.

Deroche, M. L. D., & Culling, J. F. (2011). Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation. *The Journal of the Acoustical Society of America*, *130*, 2855–2865. doi:10.1121/1.3643812.

Deroche, M. L. D., & Culling, J. F. (2013). Voice segregation by difference in fundamental frequency: Effect of masker type. *The Journal of the Acoustical Society of America*, *134*, EL465–EL470. doi:10.1121/1.4826152.

Deroche, M. L. D., & Gracco, V. L. (2019). Segregation of voices with single or double fundamental frequencies. *The Journal of the Acoustical Society of America*, *145*, 847–857. doi:10.1121/1.5090107.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *115*, 2246–2256. doi:10.1121/1.1689343.

Gardner, W. G., & Martin, K. D. (1995). HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, *97*, 3907–3908. doi:10.1121/1.412407.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*, 103–138. doi:10.1016/0378-5955(90)90170-T.

Green, T., & Rosen, S. (2013). Phase effects on the masking of speech by harmonic complexes: Variations with level. *The Journal of the Acoustical Society of America*, *134*, 2876–2883. doi:10.1121/1.4820899.

Jelfs, S., Culling, J. F., & Lavandier, M. (2011). Revision and validation of a binaural model for speech intelligibility in noise. *Hearing Research*, *275*, 96–104. doi:10.1016/j.heares.2010.12.005.

Kidd, G., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America*, *124*, 3793–3802. doi:10.1121/1.2998980.

Kidd, G., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., & Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America*, *140*, 132–144. doi:10.1121/1.4954748.

Lavandier, M., & Best, V. (2020). Modeling Binaural Speech Understanding in Complex Situations. In J. Blauert, & J. Braasch (Eds.), *The Technology of Binaural Understanding* Modern Acoustics and Signal Processing (pp. 547–578). Cham: Springer International Publishing. doi:10.1007/978-3-030-00386-9_19.

Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., & Makin, S. J. (2012). Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. *The Journal of the Acoustical Society of America*, *131*, 218–231. doi:10.1121/1.3662075.

Lavandier, M., Mason, C. R., Baltzell, L. S., & Best, V. (2021). Individual differences in speech intelligibility at a cocktail party: A modeling perspective. *The Journal of the Acoustical Society of America*, *150*, 1076–1087. doi:10.1121/10.0005851.

Leclère, T., Lavandier, M., & Deroche, M. L. (2017). The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location. *Hearing Research*, *350*, 1–10. doi:10.1016/j.heares.2017.03.012.

Patterson, R., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. In *Presented to the Institute of Acoustics Speech Group on Auditory Modelling at the Royal Signal Research Establishment.* (p. 34).

Popham, S., Boebinger, D., Ellis, D. P. W., Kawahara, H., & McDermott, J. H. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature Communications*, *9*, 2122. doi:10.1038/s41467-018-04551-8.

Prud'homme, L., Lavandier, M., & Best, V. (2020). A harmonic-cancellation-based model to predict speech intelligibility against a harmonic masker. *The Journal of the Acoustical Society of America*, *148*, 3246–3254. doi:10.1121/10.0002492.

Prud'homme, L., Lavandier, M., & Best, V. (2022). A dynamic binaural harmonic-cancellation model to predict speech intelligibility against a harmonic masker varying in intonation, temporal envelope, and location. *Submitted to this special issue, under review*, .

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, *133*, 2431–2443. doi:10.1121/1.4794379.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*, 225–246. doi:10.1109/TAU.1969.1162058.

Schoof, T., & Rosen, S. (2015). High sentence predictability increases the fluctuating masker benefit. *The Journal of the Acoustical Society of America*, *138*, EL181–EL186. doi:10.1121/1.4929627.

Steinmetzger, K., & Rosen, S. (2015). The role of periodicity in perceiving speech in quiet and in background noise. *The Journal of the Acoustical Society of America*, *138*, 3586–3599. doi:10.1121/1.4936945.

Steinmetzger, K., Zaar, J., Relaño-Iborra, H., Rosen, S., & Dau, T. (2019). Predicting the effects of periodicity on the intelligibility of masked speech:

An evaluation of different modelling approaches and their limitations. *The Journal of the Acoustical Society of America*, *146*, 2562–2576. doi:10.1121/1.5129050.

Stone, M. A., Füllgrabe, C., Mackinnon, R. C., & Moore, B. C. J. (2011). The importance for speech intelligibility of random fluctuations in "steady" background noise. *The Journal of the Acoustical Society of America*, *130*, 2874–2881. doi:10.1121/1.3641371.

Vicente, T., & Lavandier, M. (2020). Further validation of a binaural model predicting speech intelligibility against envelope-modulated noises. *Hearing Research*, *390*, 107937. doi:10.1016/j.heares.2020.107937.

Vicente, T., Lavandier, M., & Buchholz, J. M. (2020). A binaural model implementing an internal noise to predict the effect of hearing impairment on speech intelligibility in non-stationary noises. *The Journal of the Acoustical Society of America*, *148*, 3305–3317. doi:10.1121/10.0002660.

Wasiuk, P. A., Lavandier, M., Buss, E., Oleson, J., & Calandruccio, L. (2020). The effect of fundamental frequency contour similarity on multi-talker listening in older and younger adults. *The Journal of the Acoustical Society of America*, *148*, 3527–3543. doi:10.1121/10.0002661.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*, 1293–1313. doi:10.3758/BF03194544.