Towards modelling energetic masking for speech intelligibility in cocktail party situations

Luna Malka Prud'Homme

N°d'ordre NNT : 2021LYSET007

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein du
Laboratoire de tribologie et de dynamique des systèmes

École Doctorale N° 162
MEGA (Mécanique, Energétique, Génie Civil et Acoustique)

Spécialité / discipline de doctorat :
Acoustique

Soutenue publiquement le 01/07/2021, par :
Luna Prud'homme

Towards modeling energetic masking for speech intelligibility in cocktail party situations

Vers la prédiction du masquage énergétique pour l'intelligibilité de la parole dans les situations de cocktail party

---

Devant le jury composé de :

Dr. Fanny Meunier	CNRS Université Côte d'Azur	Présidente
Pr. Stuart Rosen	University College London, UK	Rapporteur
Pr. Andrew Oxenham	University of Minnesota, USA	Rapporteur
Pr. John Culling	Cardiff University, UK	Examinateur
Dr. Mathieu Lavandier	Ecole Nationale des Travaux Publics de l'Etat, Lyon	Directeur de thèse
Dr. Virginia Best	Boston University, USA	Co-directrice de thèse

Acknowledgements

I would like to thank Mat and Gin for giving me the opportunity to work on this PhD project with them and for trusting that I would finish it. I would especially like to thank Mat for always being over-optimistic to counterbalance my over-pessimism and for all the interesting discussions that we had about research (and life). I would also like to thank Gin for welcoming me to Boston, and for her always relevant advices.

Merci également à toutes les personnes aux côtés de qui j'ai travaillé à l'ENTPE. I would also like to thank everyone in the lab in Boston for welcoming me into their team.

Enfin, je tiens également à remercier tous celles et ceux, amis et famille, qui m'ont soutenue et supportée durant ces trois années.

Table of contents

List of figures

List of tables

Introduction

In everyday life, there are many situations where people are faced with the challenge of listening to a speech discourse while being disturbed by competing noises: in public transportation, restaurants, classroom, open space offices... Several mechanisms that allow the auditory system to improve speech intelligibility in such conditions have been identified in the literature. Such mechanisms rely on acoustic properties of the speech target and the masking sound sources. Many studies focused on less complex stimuli and situations than what may be encountered in real life (several competing talkers and noises) in order to simplify the problem. The extent to which these results apply to real-life situations remains to be determined.

It is important to better understand the perceptual mechanisms involved to improve speech intelligibility in such situations. Speech intelligibility models provide one way to achieve that. Accurate modeling of the mechanisms used by the auditory system to improve speech intelligibility can lead to a better understanding of those mechanisms. Such models can later be used to improve building designs, as well as hearing aid designs, if they are deemed appropriate for a relevant evaluation of those designs. Although several models have been proposed in the literature, they are not yet able to predict speech intelligibility in realistic complex situations with multiple talkers. In order to aim towards creating such models, several speech characteristics need to be taken into account. In particular, one aspect that needs to be further explored is the effects associated with fundamental frequencies (F0s, the acoustic property associated with pitch) of the target and masker signals.

First, chapter I presents the issues associated with speech intelligibility among competing speech and the current state of the art concerning modeling on the effects associated with F0. Chapters II and III present the development of a speech intelligibility model with the implementation of harmonic cancellation. The model was first validated on monotone stationary harmonic complexes (chapter II) and extended to harmonic complexes varying

in F0 contour, temporal envelope or location (chapter III). In chapter IV, a behavioral experiment was conducted to examine the role of harmonic cancellation on different types of masker ranging from noise to speech. Further insights were provided by applying the model developed in previous chapters to confirm the experimental interpretations concerning the potential role of harmonic cancellation for speech intelligibility in cocktail-party situations.

CHAPTER I

State of the art

1 Introduction: the cocktail party problem

Trying to attend to a speaker in a noisy environment is a very common situation in which several factors may prevent the listeners from correctly understanding the person talking, like interfering noises and competing talkers. In this kind of situation, often referred to as the "cocktail party problem" (Cherry, 1953), the listeners are faced with many challenges. They first need to segregate the different sources, then select the source to focus on, and finally understand the information carried by that sound source. One of the factors that can affect speech intelligibility is the signal-to-noise ratio (SNR) at which the sounds are presented. A way to investigate speech intelligibility is to conduct experiments in which listeners are presented with a speech signal in the presence of a masking source and are asked to report the target sentence that they heard. A speech intelligibility measure can be obtained by counting the percent of correct words that the listeners understood. The speech reception threshold (SRT) corresponds to the SNR at which a listener is able to understand a certain proportion of the target words (often set to 50 %). A decrease in SRT will thus correspond to an increase in intelligibility.

The majority of studies in the literature concerned with masked speech intelligibility made use of noise maskers, which are well-controlled and convenient to study. However, the case of the cocktail party situation is particularly complex to investigate because the masking sources are not only noise but also competing talkers, and the acoustic properties are different in speech and noise. Parts of the speech signal are harmonic, which means that its spectrum is composed of harmonics that are equally spaced in frequency at multiples of the fundamental frequency (F0). It is composed of vowels (that are harmonic complexes) and consonants. Some of the consonants are voiced (harmonic) whereas others are unvoiced

(not harmonic). Speech also presents intonation (variation of F0 over time) and amplitude modulations. Thus, in the specific case of speech-on-speech (SOS) masking, many factors that are not present with noise come into play.

2 Energetic and informational masking

Masking is a major component in understanding the cocktail party problem. It describes the way in which interfering sources prevent listeners from understanding a target voice, and it is often quantified in terms of the SRT. Masking can be separated in two categories: energetic masking and informational masking (Brungart et al., 2001). The definitions of energetic and informational masking are still not well fixed in the literature.

2.1 Energetic masking

Energetic masking (EM) refers to masking that occurs when the target and masker signals overlap and compete at the periphery of the auditory system (Durlach et al., 2003), ie., when target and masker signals overlap in the time and frequency domains. Culling and Stone (2017) define EM as masking that can be released by low-level processes. Several mechanisms are known to provide release from EM and will be defined in section 3.

2.2 Informational masking

In cocktail party situations, and any situation where a listener is trying to understand one talker among competing talkers, masking can occur even though the target is sufficiently audible. Informational masking (IM) refers to a reduction in speech intelligibility that cannot be explained by EM. This broad definition of IM includes a wide range of factors that prevent the listener from segregating the target speech from competing voices, or from focusing on the right talker. Contrary to EM, IM is often thought to rely on more central factors and to be due to interactions between target and masker that happen beyond the peripheral auditory system (Durlach et al., 2003; Kidd and Colburn, 2017; Watson, 2005). It is often attributed to the inability of the listener to selectively attend to the target talker and overcome confusion or distraction. IM is closely linked with stimulus uncertainty and similarity between target and masker. It is usually high in conditions where the target and masker share similar characteristics, thus the introduction of differences along any dimension between target and masker can reduce the similarity and the amount of IM. For example, several factors can provide a reduction of IM in SOS masking (Kidd and Colburn, 2017), like spatial separation, differences in sex or fundamental frequency, masker time reversal or linguistic dissimilarity.

2.3 Modulation masking

Modulation masking is thought to occur when the presence of a modulated masker prevents a listener from detecting and processing the temporal fluctuations of the target speech that could be useful for speech recognition. In particular, this can happen when the target and masker share similar modulation rates (Fogerty et al., 2016). Recent studies suggested that even steady-state noise, traditionally thought to produce mainly EM, could produce modulation masking due to its random fluctuations (Stone et al., 2011, 2012).

2.4 Difficulty to classify IM and EM

As seen in the previous sections, giving clear definitions of EM and IM is not straightforward. Many studies in the literature pointed out the difficulties in defining the terms (Durlach et al., 2003; Kidd and Colburn, 2017; Watson, 2005). One of the reasons for the difficulty comes from the multiplicity of factors that come into play in cocktail party situations, and the fact that the underlying mechanisms are not always well understood. In particular, some of the cues used by the auditory system can provide a release from both EM and IM simultaneously (e.g., spatial separation, F0 difference, see section 3). It is thus complicated to separate the effects due to EM or IM and to distinguish the two phenomena.

Several studies tried to separate the EM and IM components of SOS masking. Brungart et al. (2006) proposed a method to isolate the energetic and informational components in SOS masking by using ITFS (ideal time-frequency segregation). The idea is that ITFS signals contain only the spectro-temporal regions in which the target energy is higher than the masker energy. This would approximate a perfect segregation of the two signals (using as criterion SNR=0 dB) and with the resulting loss of target energy in masker-dominated regions approximating EM. Other studies followed this method to estimate IM (Conroy et al., 2020; Kidd et al., 2016; Rennies et al., 2019). However, as explained by Conroy et al. (2020), the estimation of IM using this method depends on the resolution of the ITFS processing and a number of underlying assumptions. Arbogast et al. (2002) took a different approach and presented the target and masker in different frequency bands, so that they never overlapped. By doing so, they eliminated EM and effectively isolated the IM in different SOS situations. Another way to estimate IM is to quantify the EM component using a model. However, most speech intelligibility models that aim to predict EM have been validated with noise maskers (e.g. see Lavandier and Best, 2020, for a review of binaural speech intelligibility models). Thus they usually do not take into account all the characteristics of speech signals and as a result may not provide accurate estimates of the EM present in SOS situations.

The challenge of estimating EM and IM in SOS situations is well illustrated by the fact that different studies have produced different conclusions. Given the results of some studies, it is possible that IM might be dominant in SOS and that EM could be minor (Arbogast et al., 2002; Brungart, 2001; Brungart et al., 2001; Conroy et al., 2020; Kidd et al., 2005). Some recent studies even conclude that there is no "pure" EM component in SOS situations and that IM and modulation masking dominate (Stone and Canavan, 2016; Stone and Moore, 2014). On the other hand, it can be argued that most realistic situations deviate substantially from laboratory situations known to be high in IM (such as same-talker target and masker, co-located condition). There is usually at least one factor that can greatly reduce confusion between target and masker: spatial separation, different voices, visual cues. Furthermore, in the presence of several speech maskers, the overall masking sound will become less intelligible and cause less IM as the number of competing talker increases (Freyman et al., 2004; Simpson and Cooke, 2005). By comparing SRTs measured against speech and unintelligible vocoded speech maskers in a simulated cafeteria, Westermann and Buchholz (2015) concluded that confusion-based IM was substantially reduced by differences in location or talker voice, suggesting that IM might be minor in similar situations. The relative contributions of IM and EM in real-life situations remain unclear and depend strongly on the conditions.

3 Unmasking mechanisms

The auditory system is able to use several mechanisms to reduce masking and improve speech intelligibility in the presence of competing sound sources. Those mechanisms rely on acoustic cues such as F0 and harmonicity, binaural differences and temporal envelope fluctuations.

3.1 F0-based effects

Several studies showed improvements in speech intelligibility when the target and harmonic masker differed in F0 for different kinds of maskers: speech maskers (Bird and Darwin, 1998; Brokx and Nooteboom, 1982; Darwin et al., 2003; Deroche and Culling, 2013), harmonic complexes (Deroche and Culling, 2011, 2013; Deroche et al., 2014a; Leclère et al., 2017), double-vowel experiments (Culling and Darwin, 1993; de Cheveigné et al., 1997a, 1995; Summerfield and Culling, 1992). Brokx and Nooteboom (1982) measured the intelligibility of a monotonized target talker against a monotonized masker talker (resynthesized to have a fixed F0). They found that the percent of errors decreased with increasing F0 difference

between target and masker, except when the difference was one octave, suggesting a role of the harmonic structure in ΔF0 (differences in F0) effects. Darwin et al. (2003) found that differences in F0 and vocal tract length both improved segregation of target and masker talkers. A large number of studies in the literature on F0-based effects have investigated this subject using vowels or harmonic complexes in order to simplify the problem. Some studies used the double-vowel paradigm and showed that pairs of vowels were better identified when introducing a ΔF0 (Culling and Darwin, 1993; de Cheveigné et al., 1997a; Summerfield and Assmann, 1991). Similar results were also found with sentences masked by harmonic complexes (Deroche and Culling, 2013; Deroche et al., 2014b; Leclère et al., 2017): a difference in mean F0 between the target and a harmonic complex masker resulted in better speech recognition.

Other studies also found that harmonicity influenced speech intelligibility (de Cheveigné et al., 1995; Deroche and Culling, 2011; Deroche et al., 2014b; Popham et al., 2018; Steinmetzger and Rosen, 2015). Several mechanisms have been proposed to explain those benefits: spectral glimpsing, harmonic cancellation, segregation by F0.

Spectral glimpsing

Deroche et al. (2014b) suggested that listeners could glimpse target energy in the spectral dips of a harmonic masker, which occur between the resolved partials. They showed that SRTs were better for maskers that provided larger or more numerous spectral glimpses. This theory is supported by other studies that found similar results (Deroche et al., 2014a; Deroche and Gracco, 2019; Guest and Oxenham, 2019). Guest and Oxenham (2019) conducted an experiment with targets having a monotonous F0 contour at different F0s (80, 95, 160 and 190 Hz) against speech-shaped harmonic complex tones at those same F0s. The target F0 was at 80 Hz and the masker was 0, 3, 12 or 15 semitones above or the other way around (masker F0 at 80 Hz and target above). In both cases, the signal at the lower F0 was either presented with all its harmonics or the even harmonics were removed (in the voiced parts of speech for the target). They found that listeners benefited from removing the even harmonics of the masker even when the target and masker shared the same pitch (ΔF0 = 0 or 12 semitones). In particular, they found that a ΔF0 of one octave was beneficial when the masker F0 was one octave above the target F0 but not in the opposite condition. Spectral glimpsing could explain the effects observed in the results : there are less opportunities for spectral glimpsing in the case where the target F0 is one octave above the masker F0, in which every target harmonic is common with the masker harmonic, than the opposite case, in which there is always a target harmonic between the masker harmonics. The spectral glimpsing explanation is also consistent with the results that SRTs were better when the target F0 was 15 semitones

above the masker F0 than when it was 15 semitones below. This is similar to Deroche et al. (2014a)'s results and explanations of spectral glimpsing: the spectral dips of the masker deepen with increasing its F0, increasing the opportunities for spectral glimpsing.

Harmonic cancellation

Another mechanism relies on the idea that listeners are able to detect the harmonic structure of the masker and suppress it when its F0 is different from that of the target ("harmonic cancellation"). In a study by de Cheveigné et al. (1997b), listeners were presented pairs of vowels that were either harmonic or inharmonic. They were asked to report the vowels they heard, and each vowel in the pair was scored separately. The overall identification rate was better when the masker was harmonic, but there was no effect of the harmonicity of the target. These results suggested that harmonicity in the masker was more important than harmonicity in the target, which supported the hypothesis of harmonic cancellation over harmonic enhancement (theory according to which intelligibility is improved by taking advantage of the harmonic structure of the target). Results from other studies also support the theory of harmonic cancellation. Deroche and Culling (2011) measured SRTs for sentences against harmonic complex tones, and investigated the effect of altering the harmonicity of the target or masker using F0 modulation and reverberation. Steinmetzger and Rosen (2015) measured SRTs for target speech with different degrees of periodicity (using different types of vocoders to create the stimuli) against speech-shaped noise and harmonic complexes with dynamic F0 contours extracted from speech. Both studies found that harmonicity in the target speech was of little importance to intelligibility but that harmonicity in the masker could improve SRTs.

F0-based segregation

Most of the studies described above focused on non-speech maskers and were concerned primarily with energetic effects. There are several characteristics of speech signals that are not taken into account when using harmonic complexes, such as variation in the F0 over time (intonation), amplitude modulations in the temporal envelope, unvoiced parts, semantic content. A number of studies suggest that F0-based mechanisms might differ between speech and non-speech maskers. Leclère et al. (2017) showed that the benefit due to differences in mean F0 between a target and a harmonic masker was greatly reduced when one of the two stimuli was intonated. This could indicate that in more realistic situations of natural speech masked by natural speech, the release from EM due to differences in mean F0 could be small compared to the effects described in the literature for monotonized signals. Deroche and Culling (2013) investigated the benefit of ΔF0 for monotonized speech against

monotonized speech or monotonized harmonic complexes. They found that the improvements in intelligibility observed by introducing a ΔF0 of 0, 2 or 8 semitones was different depending on the masker. Deroche and Gracco (2019) also observed different effects when comparing speech and harmonic complex maskers. Some studies investigated F0-based effects using speech maskers but often manipulating their F0 contour. Brokx and Nooteboom (1982) used monotonized sentences. Bird and Darwin (1998) used monotonized sentences that were entirely voiced. Both studies found that introducing a ΔF0 improved speech recognition. Jackson and Moore (2013) and Assmann (1999) investigated the influence of ΔF0 between target and masker that were either both monotonized or both naturally intonated. Assmann (1999) found a benefit of introducing a ΔF0 (although it was smaller than the benefit observed by Bird and Darwin, 1998) whereas Jackson and Moore (2013) found a benefit only in the monotonized case. The main difference between the two studies was that Assmann (1999) presented simultaneous sentences spoken by the same talker and that listeners were instructed to report words from either sentence, whereas Jackson and Moore (2013) used IEEE target sentences against a fragment of running speech spoken by a different talker (of which the mean F0 was manipulated). The results from those studies seem to indicate that ΔF0 effects are reduced or even non-existent as soon as speech contains unvoiced parts (Assmann, 1999 versus Bird and Darwin, 1998) or that confusion between target and masker is reduced (Jackson and Moore, 2013 versus Assmann, 1999).

These results suggest that F0 differences may provide release from masking via different mechanisms for speech and non-speech maskers. For speech maskers, it may be that F0 differences act primarily by reducing similarity and thus reducing IM. Different-sex talkers have been shown to lead to a release from IM (Kidd and Colburn, 2017; Kidd et al., 2016). David et al. (2017) also showed that F0 differences are used by listeners to form sequential streams of syllables. Using realistic speech mixtures, Popham et al. (2018) showed that harmonic structure is useful to the grouping processes that enable listeners to solve the cocktail party problem.

3.2 Temporal dip listening

Speech interferers are not stationary but have envelopes that are modulated in amplitude. Miller and Licklider (1950) showed that speech recognition was better when a noise masker was periodically interrupted. Later, other studies also showed that listeners were able to take advantage of amplitude modulations in the masker (Beutelmann et al., 2010; Bronkhorst and Plomp, 1992; Collin and Lavandier, 2013; Cooke, 2006; Festen and Plomp, 1990; Peters et al., 1998). In the temporal dips of the masker, the SNR is higher which allows the listeners to glimpse information from the target. This is often called dip listening. Collin

and Lavandier (2013) also showed that when speech modulations applied to noise were constant, the listeners could take advantage of the predictability of the gaps. In cocktail party situations, because of the uncertainty of the temporal dips in the maskers, the listeners might not be able to fully take advantage of dip listening.

The concept of dip listening suggests that modulation in the masker envelope gives an advantage whereas modulation masking suggests that modulation in the masker can reduce speech recognition by obstructing the modulation of the target speech. Although there is evidence for both phenomena, the relative contribution of each effect on speech intelligibility still needs further investigations (Culling and Stone, 2017; Kwon and Turner, 2001).

3.3 Spatial unmasking

It is well-known that spatial separation between target and masker improves speech intelligibility (spatial release from masking, or SRM; Hawley et al., 2004; Plomp, 1976). When a source is located on the side of the listener, interaural differences are introduced: interaural level differences (ILDs) and interaural time differences (ITDs). Two mechanisms relying on those cues are known to provide a release from EM: better-ear listening and binaural unmasking.

If the target is located on one side of the listener and the masker on the other side, the ear that is on the same side as the target will receive a better SNR than the other ear. This is due to the fact that the target will be closer to this ear and the masker signal will be attenuated by the head shadow. Better-ear listening is the ability of the listeners to take advantage of the better SNR at one ear under conditions of spatial separation.

Binaural unmasking relies on interaural phase differences between target and masker which are due to ITDs. When a source is on the side of the listener, the signal will reach one ear before the other, resulting in ITDs. If the two sources are co-located, there is no phase difference, but if the sources are separated, the listeners can take advantage of the phase differences to improve detection and intelligibility. According to the equalization-cancellation (E-C) theory (Durlach, 1972), the auditory system effectively "cancels" part of the masker to improve the internal SNR.

In SOS situations, SRM can also represent a release from IM (Kidd and Colburn, 2017). Spatial separation between target and masker introduces a perceptual difference that makes them easier to segregate. Arbogast et al. (2002) showed a large effect of spatial separation on IM by testing intelligibility in conditions where the spectral overlap between target and masker (and thus EM) was greatly reduced. Spatial separation also gives more cues to the listener about the source to focus on (Kidd et al., 2008). Schoenmaker et al. (2016) suggested

that SRM comes primarily from improved segregation, with very little contribution from binaural unmasking, in the case of SOS masking.

4 Speech intelligibility models tested with harmonic maskers

A number of speech intelligibility models have been proposed in the literature (see Lavandier and Best (2020) for a recent review of binaural models). Several of these models well describe spatial unmasking or temporal dip listening. The mechanisms underlying F0-based effects are not as well understood. None of the current models have implemented any mechanism accounting for F0-based effects and most of those models are tested and validated on aperiodic noise maskers. As this thesis is oriented towards F0-based effects, only models that have been tested on harmonic maskers will be reviewed here. Steinmetzger et al. (2019) tested four speech intelligibility models on intonated harmonic complexes from the study by Steinmetzger and Rosen (2015). Those four models were chosen because they rely on different theoretical assumptions. The next sections present a rapid overview of these models.

4.1 Extended speech intelligibility index (ESII)

The ESII is an extension to the speech intelligibility index (SII, ANSI S3.5, 1997) - proposed by Rhebergen and Versfeld (2005), that was validated for stationary and envelope-modulated noise maskers. The SII takes as inputs the long-term speech and noise spectra, and the listener's hearing thresholds. SNRs are computed within frequency bands and a weighting is applied depending on the importance that each band has for speech intelligibility. The SII is an index between 0 and 1 that corresponds to the proportion of speech information available to the listener. Rhebergen and Versfeld (2005) proposed an extension to the SII that takes into account masker fluctuations. The speech (represented by stationary speech-shaped noise) and noise signals are segmented into time frames on which the SII is computed. Those values are then averaged across time frames. Rhebergen et al. (2006) later introduced a function to take into account forward masking (masking of a target by a preceding masker).

4.2 Short-time objective intelligibility measure (STOI)

The STOI is a correlation-based model proposed by Taal et al. (2011). The inputs of the model are clean and noisy speech signals. First, the silent regions (not contributing to speech intelligibility) are removed. Then, the two signals are decomposed into one-third octave bands. Intelligibility measures are computed as the correlation between the temporal envelopes of the clean and noisy speech over short time-frames of 384 ms. They are then

averaged to obtain the STOI measure that is a scalar value having a monotonic relation with intelligibility.

4.3 Multi-resolution speech-based envelope power spectrum model (mr-sEPSM)

The mr-sEPSM model (Jørgensen et al., 2013) is based on the speech-based envelope power spectrum model (sEPSM) proposed by Jørgensen and Dau (2011). The inputs of the model are the noise and noisy speech signals. They are first passed through a gammatone filterbank. Then the envelope, extracted at the output of each auditory filter, is processed by a modulation filterbank. For each modulation filter, the modulation power from the noise and noisy speech are used to compute the (S+N)NR_{env} (ratio between the noisy speech and noise in the envelope power domain) in temporal windows which depend on the center frequency of the corresponding modulation filter. The SNRenv are averaged across time frames, and combined across modulation and auditory filters and then converted to a percent-correct prediction.

4.4 Correlation-based version of mr-sEPSM ($sEPSM^{corr}$)

The $sEPSM^{corr}$ is a hybrid model using a front end with a modulation filterbank, as in Jørgensen et al. (2013), and a correlation-based back end, as in Taal et al. (2011). The inputs of the model are clean and noisy speech signals. As done by Jørgensen et al. (2013), the signals are processed by an auditory filterbank, their envelope is extracted in each auditory filter and then processed by a modulation filterbank. The envelopes are extracted for the modulation filters with a center frequency above 10 Hz, while the ones below 10 Hz are kept unchanged. The signals are then logarithmically compressed and segmented into time windows. The time segments of the clean and noisy speech obtained are cross-correlated (like in Taal et al., 2011). The correlation coefficients obtained are then integrated across time, and averaged across modulation and auditory filters. A logistic function is used to compare the correlation metric from the model to intelligibility scores.

Steinmetzger et al. (2019) proposed a modified version of the $sEPSM^{corr}$. $sEPSM^{corr}2$ is the same as $sEPSM^{corr}$ except that the part in which the envelope is extracted from the outputs of the modulation filters is replaced by a full-wave rectification. This was done to preserve the difference between periodic and aperiodic parts of speech, which was lost when using a Hilbert envelope extraction like in the previous version of the model.

4.5 Modeling speech intelligibility with harmonic complex maskers

Steinmetzger et al. (2019) tested the four models presented here to predict the data from Steinmetzger and Rosen (2015). They used 8 of the conditions tested in Steinmetzger and Rosen (2015): target speech against speech-shaped noise and intonated harmonic complexes that were stationary or 10-Hz modulated and 4 types of target vocoded speech against stationary speech-shaped noise. None of the models could predict accurately the results from Steinmetzger and Rosen (2015). In particular, all of the models underestimated the benefit due to masker harmonicity. Out of the four models, the best performance was achieved using sEPSM$^{\text{corr}}$2, which still underestimated the benefit of masker harmonicity by about 5 dB. Steinmetzger et al. (2019) suggested that the performance of the models could be improved by implementing a mechanism of enhanced stream segregation dependent on masker harmonicity.

5 Model on which the PhD project is based on

This PhD aims towards predicting speech intelligibility in the presence of concurrent speech. This work was based on a series of SNR-based speech intelligibility models first proposed by Lavandier and Culling (2010) and then revised on several occasions (Collin and Lavandier, 2013; Jelfs et al., 2011; Lavandier et al., 2012; Vicente and Lavandier, 2020). Figure I.1 decribes the structure of the model from Vicente and Lavandier (2020).

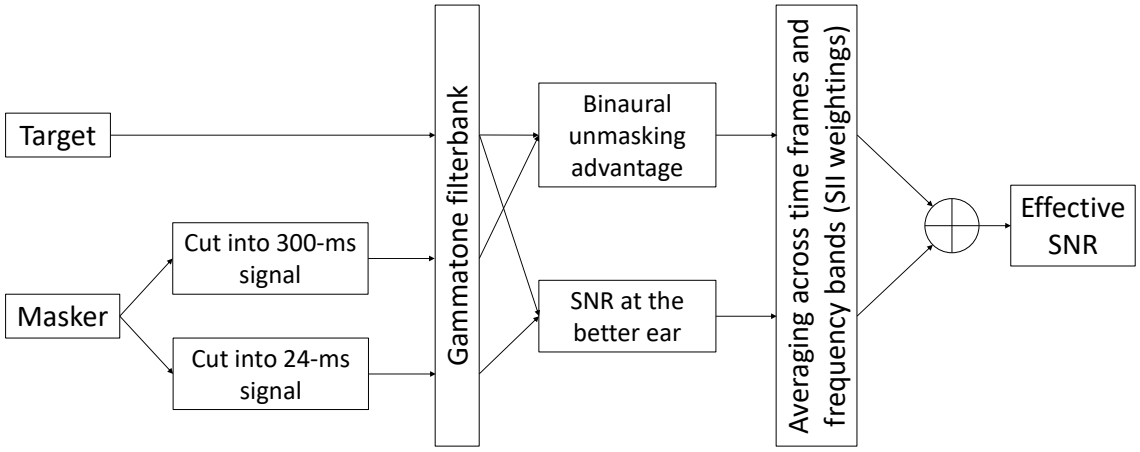

Fig. I.1 Structure of the speech intelligibility model from Vicente and Lavandier (2020).

The inputs of the model are the target and masker signals at the ears of the listener. The model computes better-ear listening and binaural unmasking (Lavandier and Culling,

2010) within time frames and frequency bands. The first versions of the model (Jelfs et al., 2011; Lavandier and Culling, 2010) could only predict intelligibility for steady-state maskers. Collin and Lavandier (2013) proposed a version of the model to take into account envelope-modulated maskers by applying it to short time frames as previously done by Rhebergen and Versfeld (2005) and Beutelmann et al. (2010). In order to consider the envelope modulation of the target speech as useful information for speech intelligibility, the model takes the average level of the target across time instead of its level within each time frame. If this was not the case, a gap in the target speech would result in a low SNR in the model, and thus induce a reduction of the predicted intelligibility. Therefore, the model uses the long-term spectrum and interaural phase of the target while it uses the short-term spectrum and interaural phase of the interferer to compute the better-ear SNR and binaural unmasking advantage. The target input to the model is an averaged target signal obtained by adding several target sentences (Vicente and Lavandier, 2020). The masker signal is cut into time frames using half-overlapping Hann windows in order to take into account temporal fluctuations of the masker (Collin and Lavandier, 2013).

The signals are then passed through a Gammatone filterbank (Patterson et al., 1987), with two filters per equivalent rectangular bandwidth. In each time frame and frequency band, the SNR is computed at each ear. The better-ear SNR is obtained by selecting the best SNR across ears band by band. A ceiling parameter was introduced into the model to prevent the SNR to tend to infinity in the temporal gaps of the masker. This ceiling corresponds to the maximum better-ear SNR that is allowed in each frequency band and time frame. This parameter was set to 20 dB (Collin and Lavandier, 2013; Vicente and Lavandier, 2020).

The binaural unmasking advantage is computed using an equation proposed by Culling et al. (2005) to estimate the binaural masking level differences (BMLDs). This formula uses the masker interaural coherence (maximum of the cross-correlation function) and the target and masker interaural phase differences (obtained by multiplying the corresponding delay by the center frequency of the band). Those parameters are obtained by cross-correlating the signals between ears. If there is no masker energy in the frequency band, the binaural unmasking advantage is set to zero.

The better-ear SNR and the binaural advantage are then integrated across frequencies using the SII-weightings (ANSI S3.5, 1997) and then averaged across time frames. The two values, computed independently, are added to obtain the effective SNR. In the latest version of the model, Vicente and Lavandier (2020) showed that the model could be optimized by using a time frame of 300 ms to compute the binaural unmasking advantage and a time frame of 24 ms to compute the better-ear SNR.

The output of the model is an effective SNR that can be compared to SRTs measured in experiments by inverting their sign so that a low SNR corresponds to a high SRT. The model can only predict relative differences between conditions but no absolute prediction of intelligibility. Therefore, a reference needs to be chosen , which is typically the average SRT across conditions (Lavandier et al., 2012).

This model was tested by Vicente and Lavandier (2020) on several data sets from the literature (Collin and Lavandier, 2013; Culling and Mansell, 2013; Ewert et al., 2017) that used a variety of target and noise masker conditions: different target speech (English, German and French), different rooms (anechoic or reverberant), different number of maskers, different masker location (azimuth, distance from listener), different types of masker modulation (steady-state, speech modulated, periodically modulated). The model performance was evaluated using mean absolute error and correlation between experimental data and model predictions. For all experiments, the correlation coefficient was between 0.85 and 0.96 and the mean error was between 0.5 and 1.4 dB.

6 Conclusion: aim of the different parts of this PhD project

The main purpose of this PhD thesis is to make progress towards predicting EM for speech intelligibility in cocktail party situations. As previous models already take into account spatial release from masking and dip listening, this PhD work focused on energetic aspects of F0-based effects. Some mechanisms have been identified in the literature on speech intelligibility (section 3) to explain release from EM due to F0, but they are not fully understood yet and to date there has been no attempt to incorporate them into a model.

First, a speech intelligibility model (based on Collin and Lavandier, 2013) was extended with an implementation of harmonic cancellation to predict speech intelligibility in the presence of stationary monotonized harmonic complexes (chapter II), then intonated, amplitude modulated or binaural harmonic complexes (chapter III). Then the role of harmonic cancellation in SOS was investigated by conducting an experiment and using the modeling results to shed light on the experimental data (chapter IV).

The broad goal was to develop a model that takes into account harmonicity-based effects to enable EM to be accurately predicted in complex mixtures of sounds. In doing so, we hoped to provide a mean by which the relative contributions of EM and IM can be confidently estimated in cocktail party mixtures. In addition, by exploring different aspects of the model, we hoped to bring new insights into the mechanisms that support speech intelligibility in the real world.

CHAPTER **II**

A harmonic-cancellation-based model to predict speech intelligibility against a harmonic masker

This chapter has been published in the Journal of Acoustical Society of America, Prud'homme, L. , Lavandier, M. and Best, V. [(2020). J. Acoust. Soc. Am. 148, 3246-3254].

1 Introduction

Speech is a complex acoustic signal that has a harmonic structure and a fundamental frequency (F0) that varies around one mean value for a particular talker. Several studies previously showed that when a speech target is masked by a competing sound that is also harmonic, F0 differences (ΔF0) between the target and masker can improve target intelligibility (Brokx and Nooteboom, 1982; Deroche et al., 2014b; Leclère et al., 2017). More broadly, it has also long been assumed that differences in voice characteristics, including F0, are critical for selectively attending to one talker in a mixture of talkers (Cherry, 1953). The mechanisms responsible for the beneficial effects of periodicity in speech-on-speech situations are still not completely understood. This is partly because, in such situations, there are two kinds of masking present, and ΔF0 may help to alleviate one or the other or both of them. Energetic masking (EM) refers to a decrease in intelligibility when the target and masker signals overlap in time and frequency such that the target becomes less audible. Informational masking (IM) refers to more central factors that can limit speech intelligibility even when the target is sufficiently audible, such as an inability to segregate the two signals or maintain selective attention to the

target speech (see review in Kidd and Colburn, 2017). IM is typically observed when the masker is very similar to the target or otherwise highly distracting.

To simplify the speech-on-speech problem, a number of studies investigated ΔF0 effects using non-speech harmonic complex maskers which cause EM but little to no IM. Several mechanisms have been proposed to explain ΔF0 benefits under such conditions: spectral glimsing (Deroche et al., 2014b) and harmonic cancellation (de Cheveigné et al., 1997a).

While several computational models are available that can predict the intelligibility of speech in various kinds of noise, to our knowledge, no model has been shown to predict effects of harmonicity and F0 segregation. Steinmetzger et al. (2019) used four existing speech intelligibility models to try to predict the data from Steinmetzger and Rosen (2015), but the best performance they obtain underestimated the effect of harmonicity by about 5 dB.

The aim of the present study was to develop an intelligibility model able to predict ΔF0 effects in the presence of a harmonic masker by extending an existing SNR-based model to include a harmonic-cancellation component. As a first step toward the prediction of ΔF0 effects, we aimed to develop a model that could accurately predict speech intelligibility against simpler stimuli than those used in Steinmetzger and Rosen (2015). We focused on stationary complex tone maskers with a fixed F0 and no amplitude modulation. This limited the factors we needed to include in this first implementation, but still represented an intermediate step between noise and more complex maskers such as speech. It also avoided possible interactions between the effects of masker periodicity and amplitude modulation (Leclère et al., 2017) that could complicate the evaluation of the model.

2 Model structure

The model presented here is based on a simplified (monaural) version of the (binaural) SNR-based model initially proposed by Collin and Lavandier (2013) and further tested by Vicente and Lavandier (2020). The inputs to the model are the target and masker stimuli at the ears of the listener. The model is composed of two parts: a basic component and a harmonic cancellation component (respectively black and grey parts in Figure II.1).

The basic component consists of four steps: (1) the signals are passed through a gammatone filterbank (Patterson et al., 1987), with two filters per equivalent rectangular bandwidth, (2) the long-term SNR is computed in each frequency band, (3) weightings are applied according to the speech intelligibility index (SII; ANSI S3.5, 1997), and (4) the SNRs are averaged across frequency bands to obtain an effective SNR.

The harmonic cancellation part of the model is implemented in parallel to the basic component. It consists of three additional steps at the front end: (1) estimation of the F0 of

the masker, (2) design of a comb filter that cancels the energy at the estimated masker F0 and its harmonics, and (3) application of this comb filter to both the target and masker signals.

Those two signals are then passed through the gammatone filterbank. The SNR is computed in each frequency band, like in the basic component of the model. SNRs with and without harmonic cancellation are computed in parallel and the higher of the two SNRs is chosen in each frequency band, with the assumption that harmonic cancellation is applied only when it is beneficial to intelligibility.

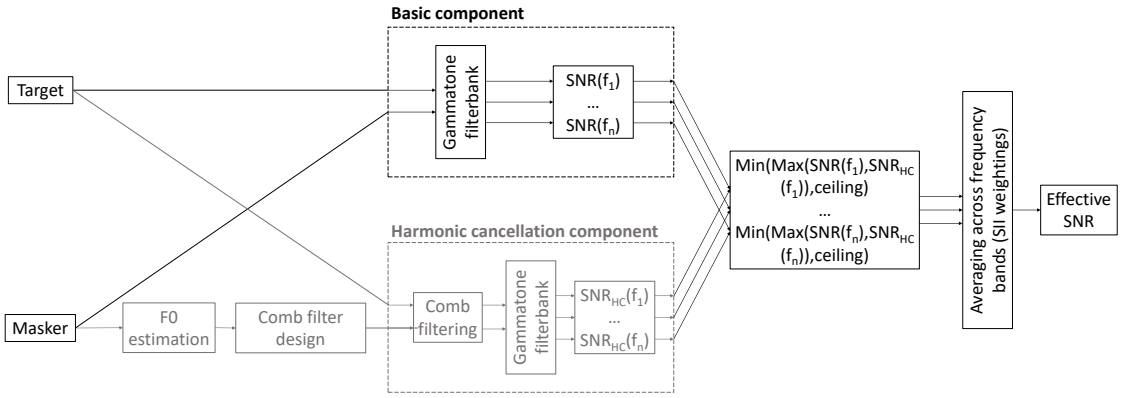

Fig. II.1 Structure of the speech intelligibility model with a harmonic cancellation component.

Four parameters are introduced at different steps of the model: a jitter in the F0 estimation, a parameter controlling the shape of the comb filter (the width of its notches), a frequency limit up to which harmonic cancellation is applied, and a ceiling to limit the highest SNR computed by the model. The rationale for the F0 jitter parameter is to simulate imperfections in the estimation of the F0 and the cancellation process. Model versions without jitter, with a fixed jitter and with a jitter increasing with the F0 of the masker were compared. Different ways to implement the comb filter were tested: one was based on the time-domain comb filter proposed by de Cheveigné (1993); the other was a frequency-domain filter in which the width and shape of the notches can be modified. A frequency limit up to which harmonic cancellation can be used was investigated, motivated by the idea that spectral components are only resolved by the auditory system within a limited range. The ceiling parameter used when computing the SNR was already present in the model of Collin and Lavandier (2013); several values were tested here to investigate potential interactions with other parameters. Table II.1 summarizes the parameters and their values tested in this study.

Table II.1 Summary of the parameters and the values tested. The final values are marked in bold.

Parameter	Values
Jitter in the masker F0	Fixed: 5-10 Hz Proportional to F0: 0; 10; 15; 20; **25**; 30 %
Shape of the comb filter	Temporal comb filter Fixed width of notches: 5-10 Hz Width proportional to F0: 0.3; 0.4; 0.5; **0.6**; 0.7F0
Upper frequency limit for harmonic cancellation	Depending on the F0 Fixed: 1000; 2000; 3000; 4000; **5000**; 6000; 7000; 8000; 9000; 10000 Hz; no limit
SNR ceiling	20; 30; **40**; 50 dB

3 Exploration of the model

3.1 Behavioral data

Deroche et al. (2014b) conducted two experiments that measured speech intelligibility against stationary harmonic complex tones with different nominal F0s and different degrees of harmonicity. These two experiments were chosen to test the proposed model that includes harmonic cancellation. Harmonic complex maskers are convenient as they allow an evaluation of the energetic effects of F0 (both spectral glimpsing and harmonic cancellation), in the absence of any significant amount of IM. Sixteen listeners performed the two experiments in the same order. In each experiment, the target stimuli were IEEE Harvard Sentences, and SRTs were measured adaptively using lists of 10 sentences.

In experiment 1, SRTs were measured in eight conditions. Maskers were harmonic or inharmonic complex tones with different F0s: 50, 100, 200 and 400 Hz. Inharmonic complex tones were created by randomly jittering (between ± F0/2) each partial from its harmonic position. For each masker type, the SRTs were measured for two conditions: frozen (the same masker was used throughout one block) or fresh (the masker was changed for each sentence). As there was no significant difference between the two conditions, the results presented here were averaged across frozen and fresh conditions. Figure II.2 (top panel, black symbols) shows the mean data reported by Deroche et al. (2014b). The key results are: (1) SRTs decrease with increasing masker F0, (2) harmonic maskers cause less masking than inharmonic maskers and (3) the difference between harmonic and inharmonic maskers decreases with increasing masker F0. To explain the first result, Deroche et al. (2014b) showed that for both harmonic and inharmonic signals, the width of spectral dips increase more than the width of spectral peaks with increasing F0, resulting in more spectral

glimpsing opportunities as the masker F0 increases. The second result suggests that there is an advantage linked to the harmonicity of the masker, which could theoretically be achieved via harmonic cancellation (or another harmonicity-based mechanism). The third result can be explained by the fact that for a given F0, spectral dips are wider for inharmonic than for harmonic maskers, and this difference increases with F0. Thus, compared to the harmonic masker, there would be more glimpsing opportunities in the inharmonic masker but less harmonic cancellation.

In experiment 2, SRTs were measured using four types of maskers with various degrees of harmonicity: a harmonic complex, an inharmonic complex created by jittering every partial of a harmonic complex, an inharmonic complex with its first two partials fixed at their harmonic positions and an inharmonic complex with its first four partials fixed at their harmonic positions. The harmonic complex had an F0 of 50, 100, 200 or 400 Hz. The inharmonic complexes were based on harmonic complexes at F0s that would create equivalent spectral glimpsing opportunities as the harmonic complex (as measured by a metric proposed by Deroche et al., 2014b). Figure II.2 (bottom panel, black symbols) shows the results from experiment 2. As in experiment 1, the SRTs decrease with increasing F0. Moreover, SRTs for the inharmonic complexes with the first two and four partials fixed are lower than for the completely inharmonic masker but higher than for the harmonic complex. This suggests that these lower partials are useful, but that all partials are needed to take full advantage of the periodicity of the masker.

3.2 Application of the model

The model described in section 2 was optimized using the data from the two experiments described in section 3.1. Stimuli from those experiments were used as inputs. Specifically, the target input was composed of 50 target sentences, concatenated to form one long target signal. The masker inputs were 160 realizations of the harmonic maskers used in the experiments; predictions were averaged across realizations. For simplicity, the masker F0 required as model input was not calculated from the signals but taken directly from the original publication of Deroche et al. (2014b).

Effective SNRs obtained with the model can be compared to SRTs by inverting their sign so that a low SNR corresponds to a high SRT. The model should only be used to predict differences in SRTs across conditions in an experiment. To compare predicted and measured SRTs, a reference needs to be chosen (Lavandier et al., 2012). In this study, the reference chosen for each experiment was the average SRT across conditions, as done by Collin and Lavandier (2013).

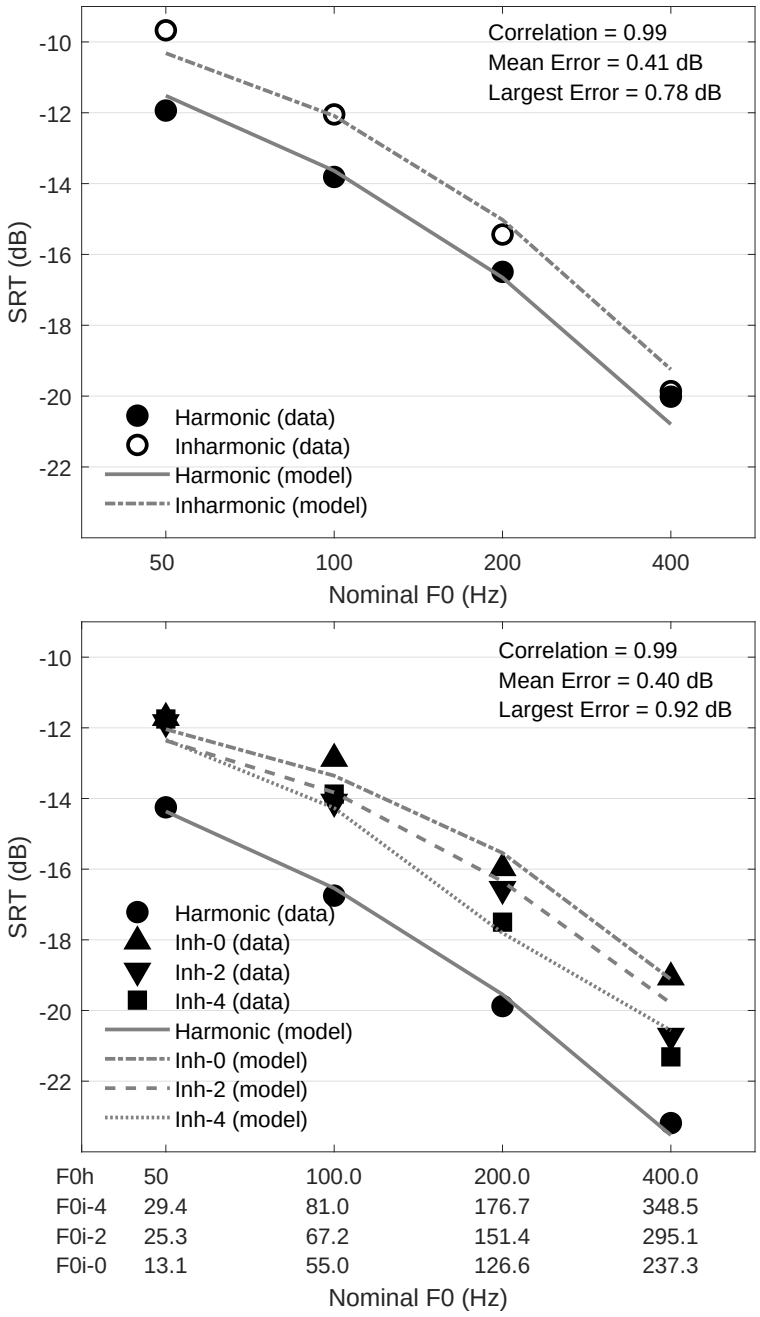

Fig. II.2 Mean SRTs measured by Deroche et al. (2014b; black symbols) and the corresponding model predictions (gray lines) for experiment 1 (top panel) and experiment 2 (bottom panel). SRTs are shown as a function of masker F0 for harmonic and inharmonic maskers. For experiment 2, Inh-0 corresponds to the completely inharmonic masker while Inh-2 and Inh-4 correspond to inharmonic maskers with respectively the first two and four partials fixed. Those signals had different nominal F0s (as listed) chosen such that the opportunity for spectral glimpsing was similar to the corresponding harmonic masker. The parameters values of the model used to generate these predictions were: jitter = 0.25F0, comb filter width = 0.6F0, frequency limit = 5000 Hz, ceiling = 40 dB.

3.3 Model predictions

Experiment 1 was first used to test the different parameters of the model and narrow down the set of parameter values (see section 3.4). The resulting parameter combinations were then tested on experiment 2 to choose an optimal combination.

Figure II.3A shows the predictions of the model for the data in experiment 1 when using only the basic component, which includes spectral glimpsing but not harmonic cancellation. The model predictions do not capture the main effect of harmonicity, confirming that spectral glimpsing alone is not sufficient to explain the results However, the model predictions do indicate that there are more spectral glimpsing opportunities as the F0 of the masker increases, especially for inharmonic maskers, confirming the explanation provided by Deroche et al. (2014b). The rest of the panels in Figure II.3 show various unsuccessful implementations of the model that will be used for illustration in section 3.4.

The final predictions generated by the fully optimized version of the model are shown in Figure II.2 (gray lines) along with the behavioral data from experiment 1 (top panel) and experiment 2 (bottom panel). The model predicts accurately the decrease in SRTs with increasing masker F0 as well as the release from masking associated with harmonicity. It also predicts the difference in SRTs between maskers with different degrees of harmonicity in experiment 2.

The performance of the model was evaluated using the mean absolute prediction error, which corresponds to the mean across conditions of the absolute difference between the behavioral data and the prediction, the largest absolute prediction error, and the Pearson's correlation between the data and the predictions. For both experiments, the Pearson's correlation between data and predictions is 0.99, the mean error is lower than 0.5 dB, and the largest error is lower than 1 dB.

As explained in section 2, four model parameters needed to be investigated in order to obtain the final predictions. After extensive evaluation of the different parameters and their combination, for which just some examples are given in section 3.4 and shown in Figure II.3, the following parameter values were selected: jitter 25 % (i.e. 0.25F0), width of the comb filter notches 0.6F0, frequency limit 5000 Hz, ceiling 40 dB. While other combinations of parameters were also successful for predicting the data from experiment 1 of Deroche et al. (2014b), they were less accurate when tested on experiment 2.

3.4 Parameter analysis

Different versions of the model and different parameters values were tested to predict the behavioral data of experiment 1 from Deroche et al. (2014b). A first analysis aimed to set a

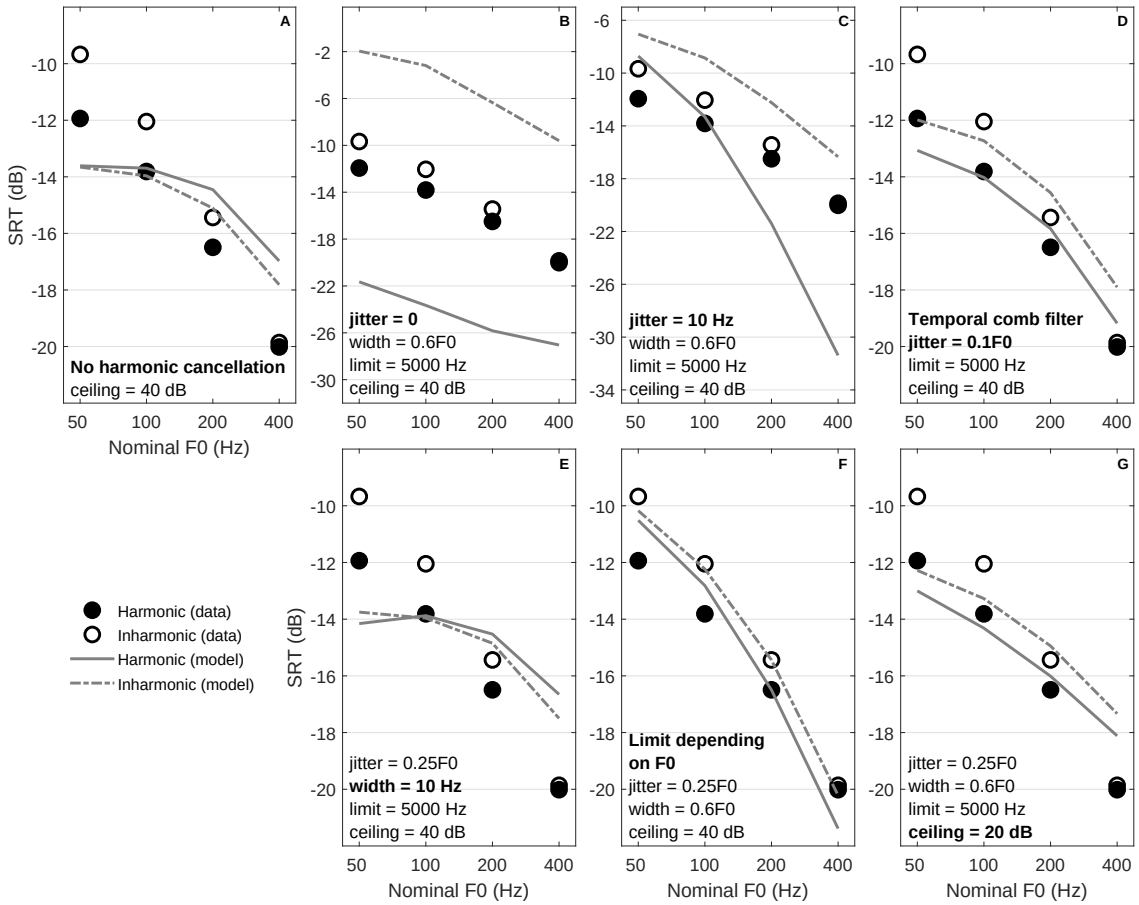

Fig. II.3 Mean SRTs measured by Deroche et al. (2014, exp. 1) and predictions of various unsuccessful implementations of the model. The parameters were set to the default values unless specified (jitter = 0.25F0, comb filter width = 0.6F0, frequency limit = 5000 Hz, ceiling = 40 dB). Panel A: using only the basic component of the model (no harmonic cancellation). Panel B: in the absence of any jitter in the F0 estimation. Panel C: using a fixed jitter in the F0 estimation. Panel D: using a temporal comb filter. Panel E: using a fixed width for the notches of the comb filter. Panel F: using a frequency limit depending on F0 (Shackleton and Carlyon, 1994; see text for details). Panel G: with a ceiling of 20 dB. Note that the ordinate is different in panels B and C.

reasonable range of values for each parameter. Then, different combinations of parameter values were tested to highlight any potential interactions between the parameters.

Introduction of an F0 jitter

Figure II.3B shows the model predictions for experiment 1 when the F0 estimation is assumed to be perfect (i.e., without introducing any jitter in this estimation). The difference between the SRTs for the harmonic and inharmonic conditions at 50 Hz predicted by the model is about 22 dB, which is 20 dB greater than the difference observed in the data (2 dB). This can be explained by the fact that the harmonic maskers are perfectly canceled by the comb filter when the F0 is perfectly estimated, whereas a physiological implementation might be less perfect.

The model was then evaluated with the introduction of a jitter in the F0 for the creation of the comb filter. The idea is to introduce a jitter in the F0 to account for the fact that the F0 might not be perfectly estimated by the brain but also for any noise in the cancellation mechanism. One solution could have been to introduce a jitter in each notch of the comb filter but introducing a jitter in the F0 was simpler and turned out to be sufficient. The magnitude of this jitter was randomly taken from a normal distribution centered at 0. In the rest of this study, the jitter parameter is defined as the standard deviation of this normal distribution.

To obtain model predictions with the jitter parameter that varies randomly from trial to trial, the model is run several times for each condition using a different realization of the stimuli and a different value of the jitter. On each of these "trials", the jitter parameter takes a different random value and produces a different prediction. These predictions are then averaged across trials to estimate the performance of the model. To determine the minimum number of trials needed to produce consistent predictions, we ran the model using 200 to 2400 trials with a jitter parameter proportional to the F0 (25%). Model performance stabilized after 800 trials, which is the number of trials used for the parameter analysis in the following sections.

Influence of the jitter

The performance of the model was explored by testing two options for defining the jitter value. The jitter could either take a fixed absolute value (in Hz) or it could be proportional to the masker F0. A selection of the results is described here.

Figure II.3C shows that when the jitter was a fixed value (10 Hz), the predicted difference between harmonic and inharmonic maskers increased dramatically with the masker F0, largely because of a very steep improvement in predicted SRTs for harmonic maskers. This is likely because 10 Hz represents a large variation at 50 Hz, but a small variation at 400

Hz, which allows for near-perfect cancellation in the latter case. This is why the difference between harmonic and inharmonic maskers is very small in the 50 Hz condition but very large in the 400 Hz condition, an effect not seen in the behavioral data.

The model was also explored using a jitter proportional to the masker F0. Several values were tested between 10 and 30% of F0. In the examples given here, this parameter was tested in combination with the parameter representing the width of the notches of the comb filter (see next section). Figure II.4 presents the mean and largest prediction errors for five jitter values between 10 and 30% of F0, and three notch width values (0.4F0, 0.5F0 and 0.6F0). Note that the performance of the model is influenced by both the jitter and the width of the notches. Overall, however, better performance is obtained when the jitter value is at least 20%. In this example, the optimal value, in which the errors are minimized or reach a plateau, is approximately 20% or 25% for the width of the notches at 0.5F0 and 0.6F0, respectively. The final model predictions using a jitter of 25% (i.e. 0.25F0) and a notch width of 0.6 F0 are shown in Figure II.2 (top panel).

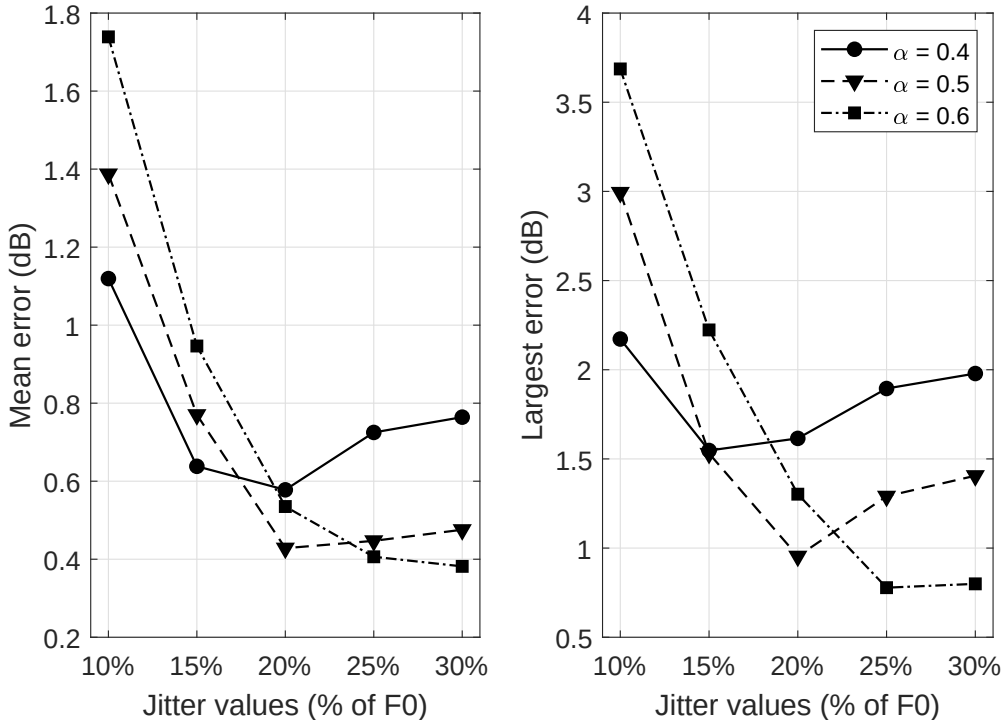

Fig. II.4 Mean and largest errors in the predictions for Deroche et al. (2014b, exp.1) as a function of the F0-dependent jitter value, for three values α of the F0-dependent width of the notches of the comb filter (frequency limit = 5000 Hz, ceiling = 40 dB).

Design of the comb filter

Two versions were tested for the comb filter that cancels the energy at F0 and its harmonics. The first option was a simple time domain comb filter proposed by de Cheveigné (1993). The impulse response of this filter is given by: $h(t) = 1/2(\delta(t) - \delta(t-T))$, where T is defined by $T = 1/F0$. Figure II.5 shows the impulse response of such a comb filter (left panel). The predictions of the model using this filter to model harmonic cancellation are displayed in Figure II.3D. The predictions are better than without any harmonic cancellation (basic component of the model, see Figure II.3A) but this model version cannot predict with high accuracy the results of Deroche et al. (2014b). In particular, the difference between harmonic and inharmonic maskers is underpredicted for lower-F0 maskers. Closer inspection of the model outputs indicated that the filter proposed by de Cheveigné (1993), when implemented with an F0 jitter, is too narrow to have a sufficient effect on the prediction.

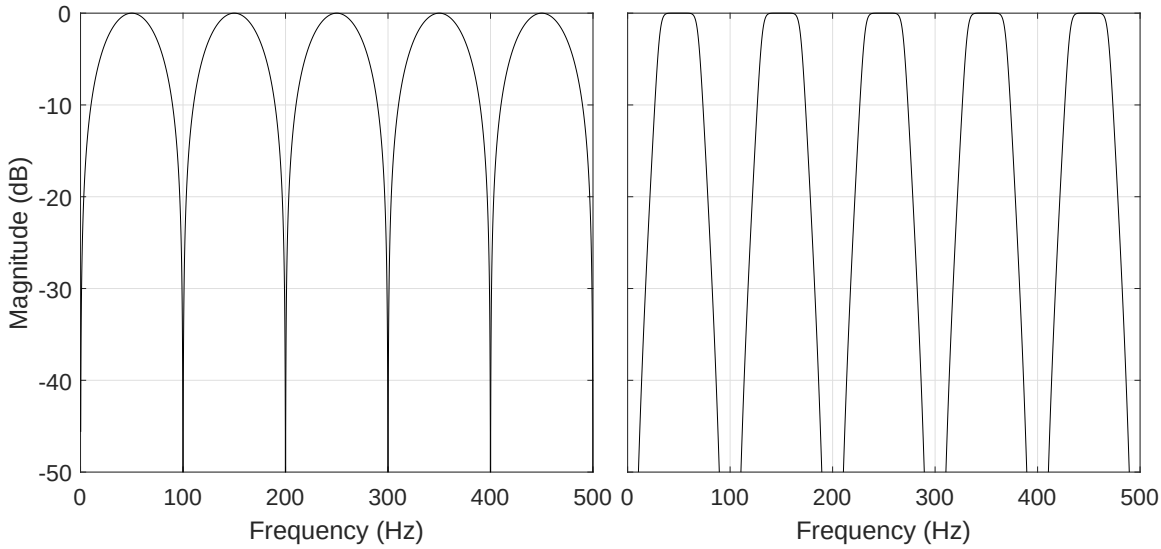

Fig. II.5 Example of the frequency response for F0 = 100 Hz of the time domain comb filter (de Cheveigné, 1993) and the comb filter used in the model with a width of notches equal to 0.6F0 (respectively left and right panels).

Another approach was considered that used an infinite impulse response (IIR) comb filter created in MATLAB using the function fdesign.comb (Figure II.5, right panel). This approach allowed us to control the width of the notches and the shape of the filter in order to cancel more of the masker energy. It was found that the width of the notches needs to be proportional to the masker F0, in order to efficiently cancel the wider peaks in the spectrum for maskers with higher F0s. Figure II.3E shows the predictions of the model when the width of the notches was fixed at 10 Hz. The model performance was significantly worse when the

width of the filter was fixed compared to when it was proportional to the masker F0 (Figure II.2). The width of the notches was then defined by width=αF0. As shown previously in Figure II.4 the best results were obtained for width of 0.5F0 or 0.6F0 (depending on the value of the jitter). The final comb filter is similar to the one proposed by de Cheveigné (1993), but it cancels more of the masker energy so that harmonic cancellation is more efficient and produces better predictions of the data.

Frequency limit

Another parameter explored was the frequency limit up to which harmonic cancellation is applied. In the frequency bands below that limit, the model is applied as explained in section 2 (choosing the maximum between the SNR from the basic component and the SNR from the harmonic cancellation component) and in the frequency bands above that limit, only the basic component is applied. The idea here is that harmonic cancellation might not be useful above a certain frequency limit (for example, once the harmonics are unresolved by the auditory system). The frequency limit could either be fixed or could depend on the masker F0, but the model performance was generally better when the frequency limit was fixed. Figure II.3F displays the model predictions when harmonic cancellation was applied only for frequency bands in the region where the harmonics of the masker would be resolved. The limit between resolved and unresolved harmonics was calculated using the definition given by Shackleton and Carlyon (1994). In this case, the frequency limit depends on the F0 of the masker. The partials are considered as resolved when fewer than two partials pass through the 10-dB bandwidth of an auditory filter and unresolved when there are more than 3.25 partials per filter. The performance of the model is not satisfactory in this case, as the difference between the SRTs of harmonic and inharmonic maskers is reduced at low F0s and increased at high F0s.

Fixed frequency limits between 1000 and 10000 Hz were tested. The model performance was poorest when the limit was below 3000 Hz. For limits of 3000 Hz and greater, the model performance was consistently good, with optimum performance at 5000 Hz (Figure II.6).

SNR ceiling

The ceiling parameter was introduced and tested in previous versions of the model that do not include harmonic cancellation (Collin and Lavandier, 2013; Vicente and Lavandier, 2020). Ceiling represents the highest value that the SNR can take in each frequency band. In Vicente and Lavandier (2020), it was fixed at 20 dB, and was necessary for accurate predictions in the presence of amplitude modulated noise maskers, where the SNR can approach infinity in the dips of the masker. While it was not clear that this parameter would be essential in

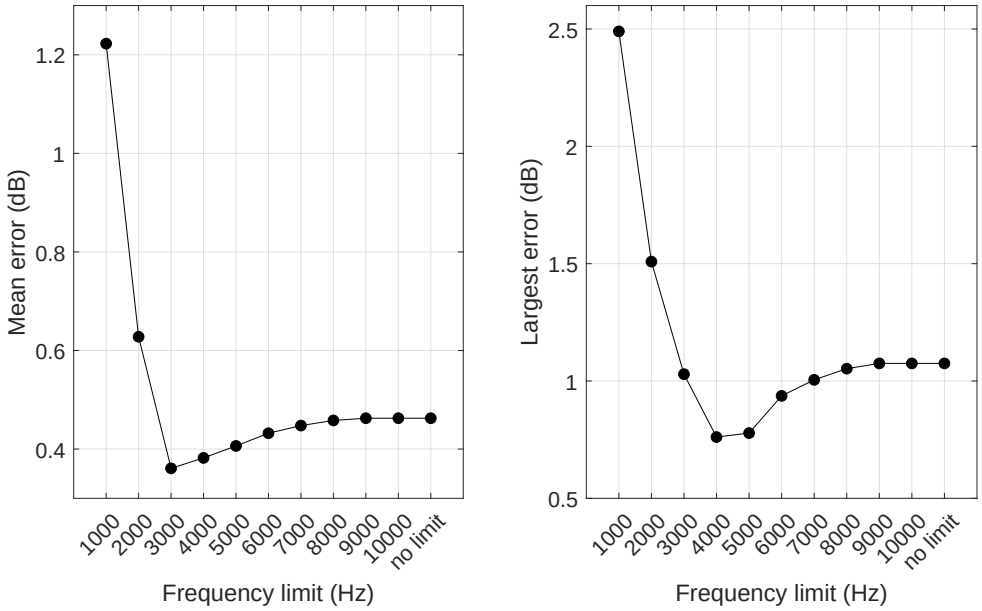

Fig. II.6 Mean and largest errors of the predictions for Deroche et al. (2014b, exp.1) as a function of the frequency limit up to which harmonic cancellation is applied. The other parameters of the model were: jitter = 0.25F0, comb filter width = 0.6F0, ceiling = 40 dB.

the present model, given that the signals are stationary, it could be important for limiting the SNR in cases where applying the comb filter greatly reduces the masker energy. After an exploration of different ceiling values (from 20 dB to 50 dB), an optimal value of 40 dB was chosen. The reason for this high value appears to be that the maskers used by Deroche et al. (2014b) were not speech shaped, but rather had a flat spectrum, which results in very high SNRs in certain frequency bands. Figure II.3G shows the predictions of the model when the ceiling was at 20 dB. This lower ceiling appears to prevent the accurate prediction of spectral glimpsing effects.

4 Discussion

The present paper describes the implementation of harmonic cancellation in an SNR-based speech intelligibility model, as well as the optimization of the new model. The model describes with a very good accuracy SRTs measured against stationary harmonic and inharmonic complexes. The mean and largest prediction errors were less than 1 dB, and similar to those reported previously for other speech intelligibility models predicting SRTs for speech in noise (Beutelmann and Brand, 2006; Collin and Lavandier, 2013; Lavandier et al., 2012). As a comparison, Steinmetzger et al. (2019) used four models to predict the masker periodicity

benefit and obtained, for the best model, a mean error of about 5 dB. Note that the present study involved simpler stimuli (with a monotonous F0) than those used by Steinmetzger et al. (2019). Note also that the proposed model still needs to be tested on data not used to define its parameters, so that its prediction power can be evaluated.

The model does not fully capture some of the effects observed in the behavioral data. In the data (experiment 1), the difference between SRTs for harmonic and inharmonic maskers is reduced with increasing masker F0, to the point that there is no difference for a masker with a F0 of 400 Hz. According to Deroche et al. (2014b), this effect is due to increased spectral glimpsing opportunities in the inharmonic compared to the harmonic masker. In the model predictions, however, SRTs are still slightly lower for the harmonic masker.

The parameter analysis that investigated whether a frequency limit for harmonic cancellation was important for the performance of the model did not result in a conclusive answer. Putting a frequency limit at 5000 Hz only resulted in a very small advantage compared to other values or running the model without a frequency limit. However, it is worth noting that the harmonic cancellation component of the model is implemented in a way that does not take into account auditory filtering. The comb filter accounting for harmonic cancellation is directly applied to the waveform of the target and masker. Introducing a frequency limit, even though its effect was limited in this experiment, could be important in future applications of the model where the use of harmonic cancellation at higher frequencies produces erroneous predictions. Further work may be needed to determine exactly what the appropriate frequency limit is. In a third experiment, Deroche et al. (2014b) showed that there was a consistent benefit due to harmonicity in the masker whether the target speech was low-pass, band-pass or high-pass filtered, which indicates that the mechanism underlying the harmonicity advantage is still active after 2535 Hz (the cut-off frequency for the high-pass filter). The results of our parameter analysis, suggesting a frequency limit close to 5000 Hz, are broadly consistent with their findings.

In the version of the model described here, harmonic cancellation was applied in each frequency band only if it improved the SNR in that band. We also tested a version of the model in which the choice between harmonic cancellation and basic component was made after averaging the SNR over all frequency bands (i.e., harmonic cancellation was applied in all or none of the frequency bands). The predictions obtained for the experimental data of Deroche et al. (2014b) were always slightly better when the choice of applying harmonic cancellation or not was made independently in each frequency band. While this is an issue that deserves further investigation, a per-channel decision seems plausible. In their discussion on the implementation of a model of harmonic cancellation, Guest and Oxenham (2019) wrote that “one simple possibility might be to selectively apply the cancellation filter to

the outputs of auditory filters that are dominated by the masker [i.e., little representation of the target periodicity is present, or the signal-to-noise ratio (SNR) is poor]. Thus, the outputs of auditory filters with a good SNR would be left unaffected while the SNR at outputs of auditory filters with unfavorable SNRs before processing might be improved by cancellation".

The next step in modeling speech intelligibility against harmonic maskers would be to take into account F0 variations over time. The maskers used here had a steady F0, which is an ideal case for harmonic cancellation, but is not representative of speech maskers, which have intonation in their F0 pattern (as well as unvoiced parts with no F0). Leclère et al. (2017) measured SRTs against both monotonized and intonated harmonic complexes, and showed that ΔF0 effects are much reduced when the masker F0 is intonated. This result implies that harmonic cancellation might be less effective when F0 varies over time and might play a smaller role (or none at all) in those situations. In order to take into account F0 variations, the model would need to operate over shorter time frames, where the F0 to be cancelled may change from frame to frame. An important step in such a modification would be to establish the appropriate duration of these time frames.

Another step forward would be to develop a model that is able to predict benefits of spatial separation and amplitude modulation in addition to harmonic cancellation. This could be approached by adding the better-ear and binaural unmasking components as implemented in Collin and Lavandier (2013). This might not be as straightforward as it seems, given that effects of better-ear listening and amplitude modulation are implemented by computing the SNR in rather short time frames (on the order of 25 ms). In the case of harmonic maskers, such time frames might be too short relative to the period of the masker in order to "see" the spectral peaks and dips in the spectrum. Thus, different time windows for different components of the model might need to be considered.

5 Conclusion

An SNR-based speech intelligibility model with an implementation of harmonic cancellation was proposed to take into account changes in energetic masking associated with F0 differences and masker harmonicity. An analysis of the four parameters introduced in the model was made, and values were chosen that optimized the performance of the model. The model was able to predict accurately the data from two experiments in which speech intelligibility was measured against maskers with different F0s and degrees of harmonicity. This work represents a critical step towards a comprehensive model that can predict speech intelligibility in more realistic situations such as those involving competing talkers.

CHAPTER III

A dynamic binaural harmonic-cancellation model to predict speech intelligibility against a harmonic masker varying in intonation, temporal envelope, and location

1 Introduction

There is currently no speech intelligibility model able to predict intelligibility in the presence of competing talkers. Some binaural models accurately predict speech intelligibility for spatially separated, amplitude modulated noise maskers (Beutelmann et al., 2010; Vicente and Lavandier, 2020) . Prud'homme et al. (2020) proposed a monaural harmonic-cancellation model that accurately predicts speech intelligibility in the presence of a stationary, monotonous, diotic harmonic masker. Speech maskers, however, present more complex characteristics than the maskers used to validate previous models. Unlike noise, speech signals are harmonic and can be characterized by their fundamental frequency (F0). Unlike monotonous harmonic complexes, speech signals contain intonation (F0 variations over time), amplitude modulations and unvoiced parts. Most real-world situations involving speech maskers also involve binaural differences. In addition, there is some evidence that these different characteristics could interact to affect the amount of masking that is ultimately observed. For example, results from Leclère et al. (2017) suggest that F0-based unmasking effects associated with harmonic complexes could be impaired by intonation. They also found that spatial release from masking (SRM), the benefit relying on binaural hearing and associated with a difference

of position for the speech target and the masker, was smaller for harmonic complexes than in previous studies with noise maskers. Two studies found that temporal dip listening, the benefit associated with envelope modulation in the masker, was larger for noise maskers than for harmonic complexes (Leclère et al., 2017; Steinmetzger and Rosen, 2015).

This chapter presents work done to extend the speech intelligibility model from Prud'homme et al. (2020) to more complex maskers containing intonation, binaural cues, and amplitude modulation. The approach taken was to try different model versions to investigate interactions between effects (F0-based effects, SRM, temporal dip listening) with these more complex maskers. This approach allowed us to estimate the influence of the different mechanisms on the energetic masking of speech. In this chapter, modeling is not only used for prediction purposes, but also to confirm hypotheses suggested by experimental results.

Our extended goal is to predict speech intelligibility for cocktail-party situations. As a step towards that goal, different model versions — variations of models from Vicente and Lavandier (2020) and Prud'homme et al. (2020) — were tested on the stimuli from Leclère et al. (2017), which are harmonic complexes varying in their F0 contour, spatial location or amplitude modulation. Those maskers are more complex than those tested by Prud'homme et al. (2020) (which were stationary, monotonous harmonic complexes from Deroche et al., 2014b), while still being simpler than natural speech maskers. Thus, they represent a step between stationary harmonic complexes and speech, allowing us to investigate how the model versions handle intonation, amplitude modulation, and binaural differences, while avoiding for now the other complex issues of unvoiced parts and informational masking.

2 Behavioral data

Leclère et al. (2017) investigated the potential interaction between F0-based effects and SRM (experiment 1) or temporal dip listening (experiment 2). They measured speech reception thresholds (SRTs, the signal-to-noise ratios, SNRs, for 50% target intelligibility) for target sentences spoken by a male voice in French. The mean F0 of the target was always fixed at 117 Hz. They tested intonated target sentences (no modification of the F0 contour of the sentence) or monotonized sentences (F0 fixed at 117 Hz). Because our model is insensitive to this difference between targets (it is operating only on the mean target spectrum, see section 3 on the model structure), we focus here on the results with the naturally intonated target. The maskers were harmonic complexes with partials in random phase. They had the same average long-term excitation pattern as the target sentences. Those speech-shaped "buzzes" were either monotonized (fixed F0 over time) or intonated (using the F0 contour of two concatenated sentences from the target speech material randomly selected on each

trial, always different from the target sentence). Their mean F0 was either equal to the target mean F0, 117 Hz, or 3 semitones above, 139 Hz, leading to a difference in mean F0 (ΔF0) of 22 Hz or 3 semitones. The target was always presented 30° to the right of the listener using anechoic HRTFs (Gardner and Martin, 1994). In experiment 1, the masker always had a stationary envelope and was tested in two spatial conditions: co-located or separated from the target (masker 30° to the left of the listener). Experiment 1 had eight conditions: 2 spatial conditions x 2 ΔF0 x 2 F0 contours. In experiment 2, all stimuli were presented diotically but the masker had either a stationary envelope or the modulated envelope of a single voice (extracted from two concatenated sentences from the target speech material, randomly selected on each trial, always different from the target sentence). Experiment 2 had eight conditions: 2 amplitude modulation x 2 ΔF0 x 2 F0 contours. The results of experiments 1 and 2 from Leclère et al. (2017) are presented in figure III.2 and figure III.3, respectively. The main results of experiment 1 were: (1) monotonized buzzes produced lower SRTs than intonated buzzes, (2) there was a ΔF0 benefit for intonated buzzes but not for monotonized buzzes, (3) the SRTs were always lower in the separated condition. The SRM was about 6 dB. While the authors noted that this amount of SRM was in good agreement with previous studies, it seems that most studies using noise maskers found SRM to be higher than 6 dB for this amount of spatial separation (60°), usually between 8 and 10 dB (Culling and Lavandier, 2021). The main results of experiment 2 were: (1) monotonized buzzes produced lower SRTs than intonated buzzes, (2) there was a ΔF0 benefit for both intonated and monotonized buzzes, (3) amplitude modulated buzzes produced lower SRTs only in the intonated condition.

3 Models

3.1 Original models

All the models tested in the present study were different variations of the models from Vicente and Lavandier (2020) and from Prud'homme et al. (2020). Both models were originally based on the model proposed by Collin and Lavandier (2013), so they have the same structure: (1) target and masker signals are passed through a gammatone filterbank with two filters per equivalent rectangular bandwidth, (2) the SNR is computed in each frequency band, (3) weightings are applied according to the SII, (4) the weighted SNRs are summed across frequency bands.

The model from Vicente and Lavandier (2020) is able to predict accurately binaural speech intelligibility for noise maskers with amplitude modulations. Amplitude modulations

in the masker were taken into account by segmenting the masker signal into short time frames using half-overlapping Hann windows. In each frequency band the binaural unmasking advantage is computed using an equation proposed by Culling et al. (2005) to estimate the binaural masking level differences (BMLDs). In parallel, the better-ear SNR is obtained by selecting the best SNR across ears band by band. The binaural unmasking advantage and the better-ear SNR are then integrated across frequencies using the SII-weightings (ANSI S3.5, 1997) and then averaged across time frames. The two values, computed independently, are added to obtain the effective SNR. A ceiling, which corresponds to the maximum better-ear SNR that is allowed in each frequency band and time frame, was introduced into the model to prevent the SNR to tend to infinity in the temporal gaps of the masker. Vicente and Lavandier (2020) showed that the model was optimized by using a time frame of 300 ms to compute the binaural unmasking advantage, a time frame of 24 ms to compute the better-ear SNR, and a ceiling of 20 dB.

Prud'homme et al. (2020)'s model is able to predict speech intelligibility for monotonous, stationary, diotic harmonic complexes. The harmonic cancellation component is implemented by filtering the target and masker signals with a comb filter that removes energy at the F0 of the masker and its harmonics. In each frequency band, harmonic cancellation is applied only if it improves the SNR in that band. Four parameters were fixed by Prud'homme et al. (2020): a jitter in the estimation of the F0 (0.25F0), the width of the notches of the comb filter (0.6F0), a SNR ceiling (40 dB), and a frequency limit up to which harmonic cancellation is applied (5000 Hz).

3.2 Tested models

Four model versions were tested here :

- Model 1: non-stationary binaural model without harmonic cancellation (Vicente and Lavandier, 2020).
- Model 2: stationary diotic model with harmonic cancellation (Prud'homme et al., 2020).
- Model 3: non-stationary, binaural model with harmonic cancellation.
- Model 4: non-stationary, binaural model with harmonic cancellation in which binaural unmasking and harmonic cancellation are mutually exclusive. Binaural unmasking is not computed when harmonic cancellation is applied.

Model 3 is a hybrid of the two existing models (1 and 2). The structure of the model is presented in figure III.1. The masker is segmented into time frames using half-overlapping

Hann windows like in model 1. The mean F0 in each time frame is computed using PRAAT PSOLA (Boersma and Weenink, 2018). If an F0 is found, the harmonic cancellation component is used (the SNRs are computed after the signals have been comb-filtered as in model 2). In the case of buzz maskers, there was always an F0 so this was not an issue but this option was still added to the model so the model can be generalized to be applied to stimuli with unvoiced parts. The mean F0 across the time frame is used as input of the comb filter. For simplicity, the same time frame duration is used for the computation of all mechanisms: harmonic cancellation, better-ear listening and binaural unmasking (unlike in model 1). The parameters set by Prud'homme et al. (2020) for the harmonic cancellation mechanism were kept unchanged: jitter (0.25F0), width of the notches of the filter (0.6F0), frequency limit for harmonic cancellation (5000 Hz). The ceiling is set to 40 dB as per Prud'homme et al. (2020).

Compared to model 2, a new parameter, the time frame duration, was introduced in the models 3. The rationale behind segmenting the masker into time frames is to account for the amplitude modulations in the masker (as in model 1) and to account for the F0 variations over time in the case of intonated maskers. In figure III.1, the bottom box ("If there is an F0") is equivalent to model 2 operating in time frames on the two ears and taking the best of the two SNRs, and adding the binaural unmasking component. The upper box ("If there is no F0"), is equivalent to model 1 using the same time frame for both binaural unmasking and better-ear listening with a ceiling at 40 dB instead of 20 dB.

Model 4 is a modified version of model 3. It has the same structure, except that when the harmonic cancellation component is on (ie., when there is an F0), binaural unmasking is not computed (gray part in figure III.1). This hypothesis suggests that the mechanisms of harmonic cancellation and binaural unmasking are mutually exclusive.

The new parameter (time frame duration) had most influence for the experiment 2 that involved fluctuating maskers, thus different time frame durations were tested for this experiment. This allowed us to define a range of acceptable values for the time frame duration that gave good predictions for experiment 2. Some of these values were then tested for experiment 1 to find the final value that provided the best fit to the experimental data for both experiments.

3.3 Implementation and evaluation of the predictions

For all model versions, the model input for the target was an averaged target signal created by adding 120 target sentences, truncated to the duration of the shortest sentence. For the harmonic cancellation models (models 2, 3 and 4), the predictions were computed using 800 trials, as done by Prud'homme et al. (2020).

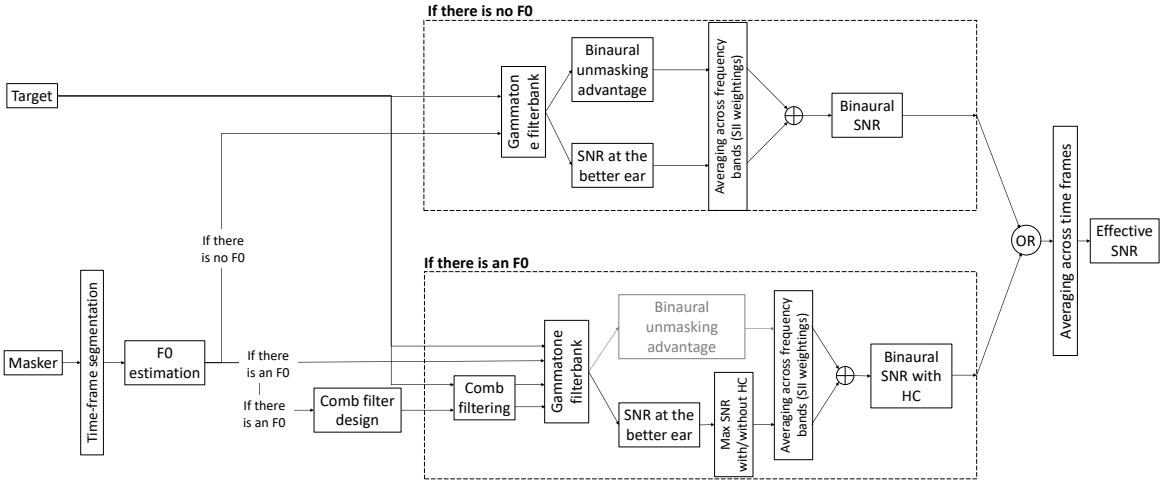

Fig. III.1 Structure of the binaural speech intelligibility model with harmonic cancellation (model 3)

The performance of the model was evaluated using the mean absolute prediction error, which corresponds to the mean across conditions of the absolute difference between the behavioral data and the prediction, the largest absolute prediction error, and the Pearson's correlation between the data and the predictions.

4 Results

4.1 Previous models (models 1 and 2)

Figure III.2(A) presents the predictions from model 1 for experiment 1. The model predicts a difference in SRT between the co-located and separated conditions (SRM) of about 10 dB, which is about 4 dB higher than the SRM observed in the data. As expected, this model does not predict any of the F0 effects: it does not predict the difference between monotonized and intonated maskers, nor the ΔF0 benefit observed in the experiment.

Figure III.2(B) presents the predictions from model 2 for experiment 1. This model is monaural and cannot predict the difference between co-located and separated conditions, thus only the co-located conditions are presented here. The predictions were computed using the right ear only. Compared to model 1, which does not implement harmonic cancellation, this model predicts the difference between intonated and monotonized buzzes with good accuracy (mean error = 0.8 dB, largest error = 1.3 dB). The model also predicts a SRT difference associated with ΔF0, but it is larger than that observed in the data for monotonized maskers

(1 dB compared to 0.8 dB in the data) and smaller than that observed in the data for intonated maskers (-0.1 dB compared to 0.8 dB in the data).

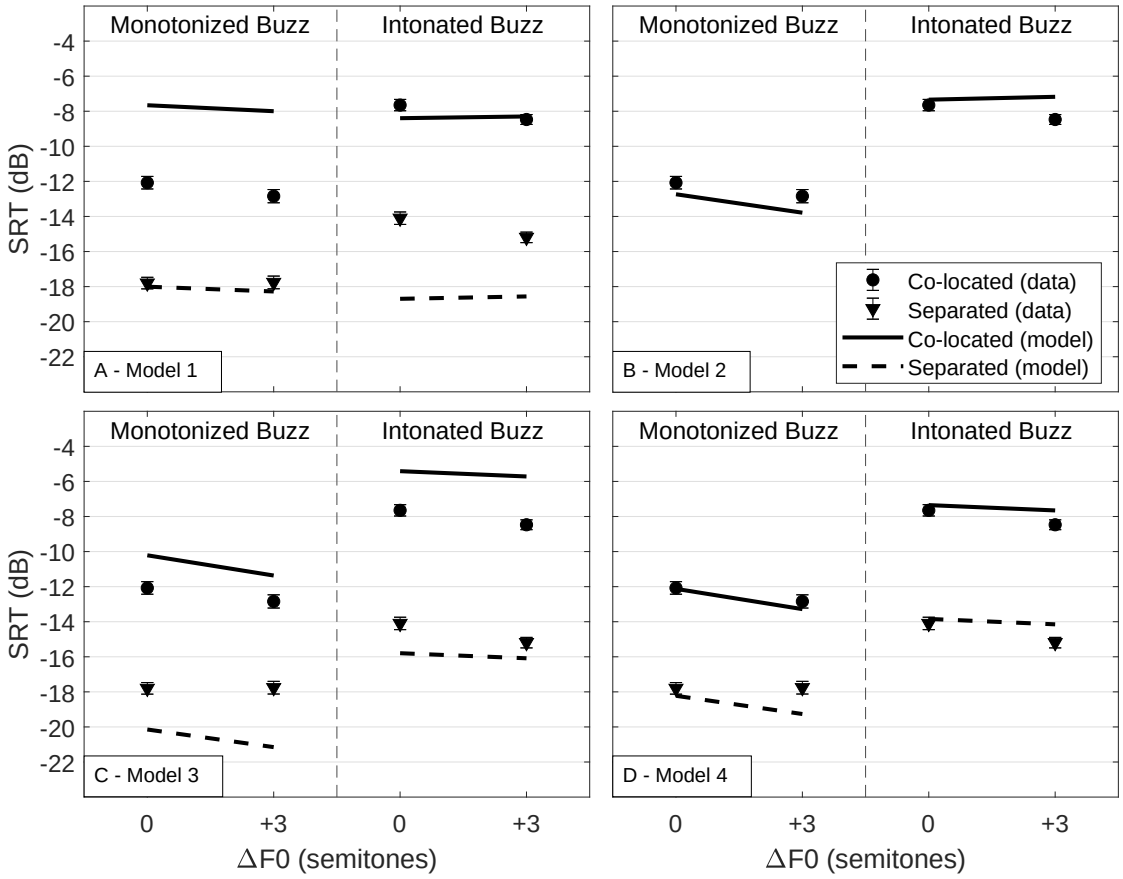

Fig. III.2 Mean SRTs measured by Leclère et al. (2017) in experiment 1 for stationary co-located and separated buzzes, with the corresponding model predictions using: (A) model 1: a binaural model without harmonic cancellation (Vicente and Lavandier, 2020), (B) model 2: a monaural model with harmonic cancellation (Prud'homme et al., 2020), for which only the co-located condition was considered, (C) model 3: a binaural model with harmonic cancellation, (D) model 4: a binaural model with harmonic cancellation without binaural unmasking. A time frame of 300 ms was used for model 3 and 4.

Figure III.3(A) presents the predictions for experiment 2 using model 1. As for experiment 1, it does not predict the difference between intonated and monotonized buzzes, the ΔF0 effect, nor the interaction of F0 effects with other effects (dip listening here). Most dramatically, this model predicts a dip listening advantage of about 4 dB for all masker types, whereas the data showed no advantage for the monotonized buzzes and only 1 dB for intonated buzzes.

Figure III.3(B) presents the predictions for experiment 2 using model 2. Contrary to model 1, model 2 predicts an opposite effect to dip listening: the predicted SRTs are lower for stationary than for modulated maskers. In particular, this difference was larger for monotonized buzzes. This could be due to the fact that harmonic cancellation was probably less effective on modulated buzzes as it does not have an effect in the temporal dips of the masker.

The two existing models cannot predict the data from Leclère et al. (2017). Model 1 failed to predict the F0-based effects, while model 2 failed to predict spatial unmasking and temporal dip listening.

4.2 New tested models (models 3 and 4)

Figure III.3(C) presents the predictions of experiment 2 using model 3. As the target and maskers were not spatially separated in this experiment, model 3 and 4 gave the same predictions. Several time frame durations were tested: from 100 to 900 ms. For clarity purposes, figure III.3(C) presents only the predictions obtained with duration of 100, 300 and 500 ms. Figure III.4 presents the mean and maximum prediction errors as a function of the time frame duration. The predictions were worse for shorter time frames (below 300 ms) and figure III.3(C) indicates that the difference between monotonized and intonated buzzes is greatly underestimated then (light gray lines). The errors were lowest for time frame durations between 300 and 800 ms (largest error < 1.5 dB, mean error < 0.7 dB). Figure III.3(C) shows that increasing the time frame duration reduced the predicted dip listening, to the point that the SRT predicted for stationary monotonized buzzes was higher than those for modulated monotonized buzzes (for durations above 500 ms). This trend was also found in the data although the difference between those two conditions was not significant. The model predicts the difference between intonated and monotonized maskers reasonably well for all time frames longer than 300 ms, although it was best for 300 ms. However, the ΔF0 effect is underestimated by the model for intonated buzzes.

Figure III.4 also presents the mean and largest prediction errors of model 4 for experiment 1 for some time frame durations. Contrary to experiment 2, increasing the time frame duration increases the errors in predictions. This is due to the fact that increasing the time frame duration increases the predicted difference between intonated and monotonized buzzes. With shorter time frames, the difference between ΔF0 = 0 and ΔF0 = 3 slightly increased for the intonated case. All the trends concerning the effect of the time frame duration were the same for model 3 and 4. Overall, the time frame duration of 300 ms resulted in good predictions for the two experiments (mean error < 0.6 dB and largest error < 1.5 dB for both experiments).

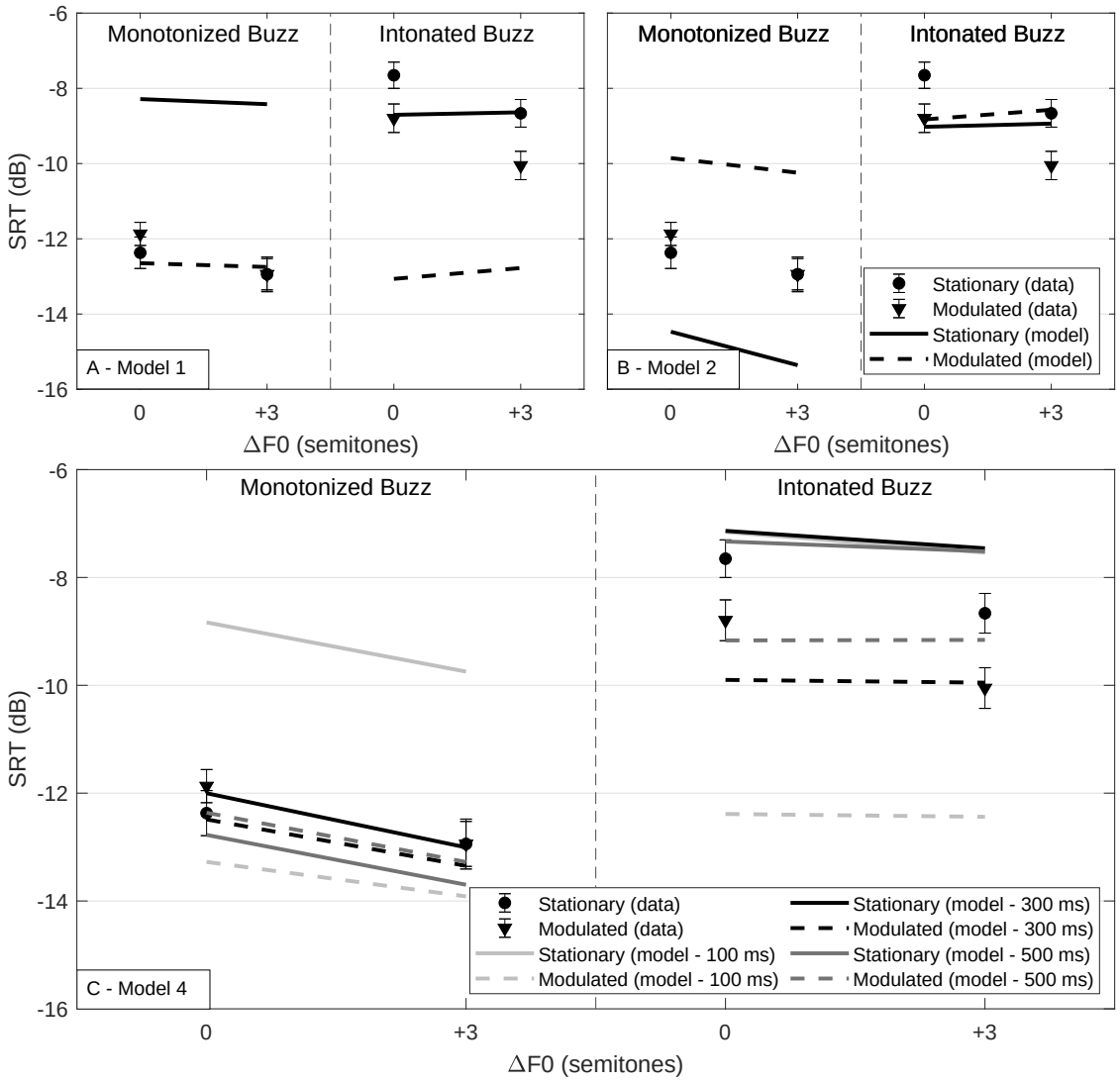

Fig. III.3 Mean SRTs measured by Leclère et al. (2017) in experiment 2 for diotic stationary and modulated buzzes, with the corresponding model predictions using: (A) model 1, a binaural model without harmonic cancellation (Vicente and Lavandier, 2020), (B) model 2, a monaural stationary model with harmonic cancellation (Prud'homme et al., 2020), (C) model 4, a binaural model with harmonic cancellation without binaural unmasking using time frame durations of 100, 300 or 500 ms. The predictions of model 3 are similar to those of model 4 in these diotic conditions.

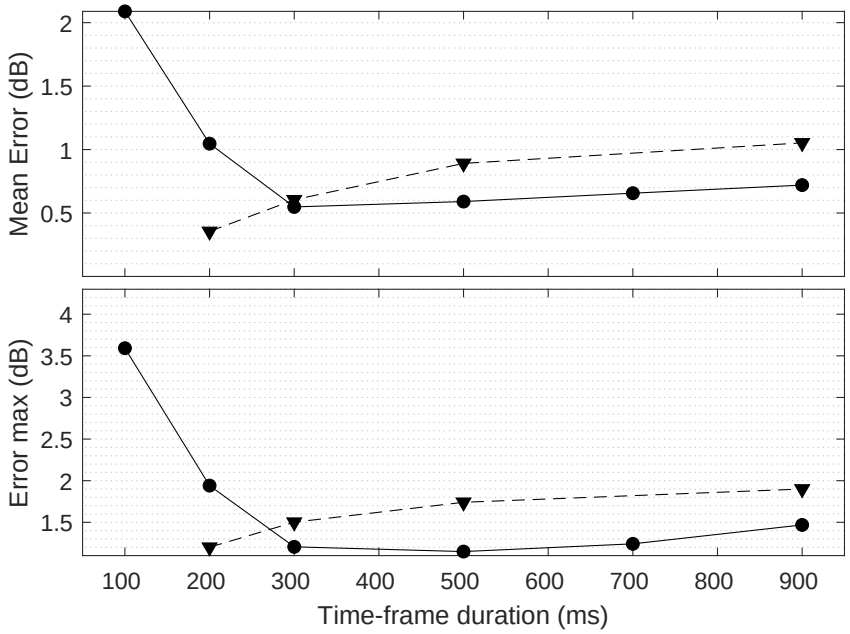

Fig. III.4 Mean and largest prediction errors as a function of the time frame duration used for the predictions of experiment 2 (circles) and experiment 1 (triangles) with model 4.

Figure III.2(C and D) presents the predictions from the non-stationary binaural models with harmonic cancellation (models 3 and 4) using a time frame length of 300 ms. These models accurately predicts the F0-based effects, as the monaural model with harmonic cancellation (model 2) does. The SRM predicted by model 3 is almost as big as the one predicted by model 1 (around 10 dB), which is larger than the observed effect (around 6 dB). The SRM predicted by model 4 was close to the SRM observed in the data (6 dB in both the prediction and the data), even though it resulted solely from better-ear effects with no contribution from binaural unmasking. Several values of the ceiling parameter were also tested (20, 30 and 40 dB) for models 3 and 4, but reducing the ceiling resulted in poorer predictions as the harmonic cancellation was less effective.

The model that provided the best fit to both experiments was model 4, using a time frame duration of 300 ms (mean error = 0.60 dB, largest error = 1.50 dB, correlation = 0.99 for experiment 1 and mean error = 0.55 dB, largest error = 1.21 dB, correlation = 0.96 for experiment 2).

5 Discussion

5.1 Intonation

As shown by the predictions from both experiments, the implementation of harmonic cancellation in the model is necessary to predict the differences between intonated and monotonized harmonic maskers. The model 1, without harmonic cancellation, completely fails (figure III.2(A) and figure III.3(A)). These results support the idea that harmonic cancellation (or a related mechanism) plays a role in the unmasking of speech in the presence of harmonic maskers, and that this mechanism operates most effectively when the F0 does not vary over time. Specifically, it appears that a monotonous F0 is easier to "cancel" than a variable F0. Leclère et al. (2017) hypothesized that the sluggishness of the mechanism, becoming less effective as the F0 varies over time, could explain the difference between intonated and monotonized buzzes. In the model, this is represented by the fact that the comb filter uses the mean F0 across the time frame. In the intonated case, the comb filter will thus be less efficient at cancelling the masker energy, because it is based in an approximated masker F0, compared to the monotonized case where the F0 is constant across the time frame. Varying the time frame duration in the model provided a test of this hypothesis: the shorter the time frame is, the smaller the approximation on the masker F0 and the more effective the harmonic cancellation will be. As shown in the results (figure III.2(B)), even the stationary harmonic cancellation model from Prud'homme et al. (2020) gave a good prediction of the effect of intonation (even if this model completely failed to predict the effects of envelope modulations, see figure III.3(B)). However, making the harmonic cancellation model non-stationary and changing the time frame duration has an influence on the predictions. A time frame of 300 ms gave slightly better predictions of the effect of intonation than the stationary model (the average error for the prediction of the effect of intonation across the two experiments are 1.8 dB for model 2 and 0.9 dB for the 300-ms model 4). The non-stationary model allows the F0 of the intonated masker to be followed more accurately, thus making the harmonic cancellation more efficient. If the time frame is too short, however, the predictions of the model are worse: the predicted SRTs are too low for intonated maskers compared to monotonized maskers (see figure III.3(C), light gray lines for the model with 100-ms time frame duration). It appears that the time frame needs to be long enough to apply harmonic cancellation effectively, but short enough so that it follows to some extent the F0 of the intonated masker.

5.2 ΔF0 benefit

The binaural model without harmonic cancellation does not predict the ΔF0 benefit (figure III.2(A) and figure III.3(A)). Leclère et al. (2017) suggested that F0-based effects observed in their data could be due to spectral glimpsing or harmonic cancellation. It is not possible to come to a conclusion to which mechanism is at play from the data only. Given that model 1 was not able to give accurate predictions of the data, the present results seem to point towards a role of harmonic cancellation, or at least that spectral glimpsing seems unable to explain the benefit due to an F0 difference between target and masker, confirming the same observation made by Prud'homme et al. (2020) while considering monotonized buzzes and different data sets. If this were the case, the model without harmonic cancellation should be able to predict the difference as the SNR is computed by frequency band, so it should be able to account at least to some extent for the effect of spectral glimpsing.

For the harmonic cancellation models, the predicted ΔF0 benefit is still small: around 1 dB for monotonized maskers and 0.3 dB for intonated maskers. In comparison, this benefit was between 0 and 1 dB for monotonized maskers and between 0.8 and 1.2 dB for intonated maskers in the data, while the model without harmonic cancellation predicted a benefit of 0.3 and -0.1 dB, respectively. The harmonic cancellation models predict a ΔF0 effect that is larger than observed for monotonized buzzes in experiment 1, in which the effect was not significant. However, Leclère et al. (2017) suggested that the non-significant effect in the data might be due to a floor effect as the SRTs were very low (around -18 dB). In experiment 2, the ΔF0 benefit observed was between 0.5 and 1 dB for the monotonized buzzes, and the model predicted a benefit of about 1 dB. To summarize, model 3 and 4 always predict a ΔF0 benefit of about 1 dB for monotonized buzzes, which is slightly larger than the effect observed in the experiment.

For intonated buzzes, the model predicts a ΔF0 benefit smaller than observed in the data (around 0.3 dB compared to 0.8 to 1.2 dB in the data). In the model, harmonic cancellation is less effective for intonated buzzes than for monotonized buzzes, which is why the model can account for the effect of intonation. This also results in a smaller ΔF0 benefit in the intonated case. As it is the case with the effect of intonation, the prediction of the ΔF0 effect is also influenced by the time frame duration. When the time frame is longer, the predicted ΔF0 benefit is smaller for intonated buzzes.

Overall, even though the model does not perfectly predict the effect of the ΔF0, this effect is quite small (less than 1.5 dB in all conditions) and the resulting errors are also relatively small.

5.3 Spatial separation

The SRM observed in experiment 1 of Leclère et al. (2017) for buzz maskers was slightly smaller than what was found in previous studies for noise maskers. There is no clear explanation for this fact. It is possible that the buzz produces less masking than noise (due to its spectral dips) and as such there is less room for SRM. Another explanation is that a mechanism linked to harmonicity (possibly harmonic cancellation) already lowered the SRTs, giving less opportunities for SRM. The binaural model without harmonic cancellation (model 1) predicts a SRM of about 10 dB for both intonated and monotonized maskers. The model predicts the same SRM as it would for noise maskers. This would suggest that the difference in SRM between noise and buzz maskers does not come from the spectral dips (taken into account in model 1) making the buzz less masking than noise. The binaural model with harmonic cancellation (model 3) also predicted a SRM larger than in the data. It is possible that masking is not additive on the whole SNR range which would be why the SRM is reduced in the data. If this were the case, the model 3 as it is now could not predict it. A model predicting psychometric functions instead of an effective SNR could be more appropriate for that. The model 4, deliberately not accounting for binaural unmasking, gives very good predictions of SRM. This result is surprising and needs further investigation. In particular, it seems worth exploring the possibility that the auditory system cannot perform harmonic cancellation and binaural unmasking at the same time within the same frequency channel. In particular, with the way that it is implemented, the model cannot fully take into account the fact that, with spectrally sparse masker like buzzes, there would be less spectrotemporal overlap between target and masker. Thus, one way to take into account the fact that binaural unmasking is a release from simultaneous masking is to make harmonic cancellation and binaural unmasking mutually exclusive.

5.4 Amplitude modulation

In experiment 2, there was no spatial separation between target and masker. The maskers were either stationary or modulated in amplitude. Previous studies comparing a stationary noise to a single-voice modulated noise masker found dip listening advantages between 4 and 12 dB (Beutelmann et al., 2010; Collin and Lavandier, 2013; Festen and Plomp, 1990; Hawley et al., 2004; Peters et al., 1998). In the experiment presented here, Leclère et al. (2017) only found a dip listening advantage for intonated buzzes (not for monotonized buzzes) and it was less than 1.5 dB. They proposed two potential explanations for this interaction. The first one was that spectral glimpsing and temporal glimpsing could not happen at the same time. The second interpretation for this interaction was that the F0 contour of the intonated buzz

allowed the listeners to anticipate the envelope fluctuations and thus make better use of the information in the dips. If this second explanation were true, then none of the models would be able to predict the difference.

The binaural model 1 without harmonic cancellation predicts a 4 dB dip listening advantage for modulated buzzes. This model gives the same predictions as it would have for noise maskers in the same configuration. If, as proposed by Leclère et al. (2017), spectral glimpsing and temporal glimpsing cannot happen simultaneously, the model 1 from Vicente and Lavandier (2020) could not account for such interaction as the structure of the model does not allow to separate the two mechanisms.

With models 3 and 4 however, increasing the time frame length reduces the predicted dip listening advantage more for the monotonized maskers than for the intonated ones. This interaction between intonation and amplitude modulation is also observed in the data. One explanation is that for monotonized buzzes, contrary to noise, the modulated masker does not provide much advantage over the stationary masker because the gaps in the signal makes harmonic cancellation less efficient. These modeling results seem to be at least partly in agreement with the suggestion from Leclère et al. (2017) that the mechanisms linked to harmonicity and dip listening cannot operate at the same time. Increasing the time frame duration is one way to model this interaction. In the model, the longer time frame needed for harmonic cancellation to predict F0-based effects limits the model's ability to take full advantage of the temporal dips of the masker, thus reducing the dip listening effect.

6 Conclusion

The model proposed here gives accurate predictions of two experiments that measured SRTs for speech masked by harmonic complexes having different F0s, different F0 contours, amplitude modulation and spatial separation. A harmonic cancellation mechanism was needed in order to predict F0-based effects and predictions were best when operating the model on 300-ms time frames. However, the most successful version of the model supposes that the mechanisms of harmonic cancellation and binaural unmasking are mutually exclusive, which raises questions about the additivity of masking that deserve further investigation.

CHAPTER **IV**

Investigating the potential role of harmonic cancellation for the intelligibility of speech masked by competing speech

1 Introduction

In cocktail party scenarios, 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 harmonicity between sources. Several previous studies showed an effect of F0 differences and/or an influence of harmonicity on speech-on-speech masking, although it is not clear what mechanisms underlie these effects. One of the complicating factors is that cocktail party situations involve both energetic masking (EM) and informational masking (IM). 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 selectively attend to the target talker. Because EM and IM typically co-occur in cocktail party situations, it is difficult to determine whether F0- and harmonicity-based improvements in intelligibility are due to reductions in EM or IM or both.

Brokx and Nooteboom (1982) tested the intelligibility of monotonized target speech against monotonized masker speech. They found that the percent 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. Popham et al. (2018) investigated the role of

harmonicity in speech intelligibility. They tested speech intelligibility with target and masker that were either both natural (harmonic) speech or both inharmonic speech. The number of correct words was reduced when the target and masker speech were inharmonic. Their results suggest that harmonicity helps grouping in speech mixtures.

In order to better understand F0 effects on speech intelligibility, several studies simplified the problem by using harmonic complexes — often called buzzes — instead of speech as maskers (Deroche and Culling, 2013; Leclère et al., 2017; Steinmetzger and Rosen, 2015). These studies confirm that, even when the maskers are simple stimuli that cause primarily EM, harmonicity reduces the masking they cause. It is unclear though whether this reduction in EM translates to the more complex case of speech maskers. With speech maskers, there are certain signal characteristics to consider that are not present in harmonic complexes: the presence of unvoiced segments, intonation, and amplitude modulation. There are some indications that these factors may reduce the role of harmonicity for speech-on-speech situations (Deroche and Culling, 2013; Deroche and Gracco, 2019; Leclère et al., 2017). For example, Leclère et al. (2017) showed that the advantage due to harmonicity was reduced for intonated buzzes compared to monotonized buzzes. Deroche and 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 aim of the present study was to investigate whether harmonicity (and harmonic cancellation in particular) plays a role in reducing EM in speech-on-speech situations. While previous experiments tested conditions with speech, buzz or noise maskers separately, all of these maskers were tested under the same conditions and on the same subjects here.

2 Main experiment

2.1 Rationale

In the present study, speech intelligibility was measured against different types of maskers, ranging from noise to speech. By comparing SRTs across masker types, the aim was to estimate the influence of harmonicity while controlling for various other characteristics of speech. In this experiment, IM was deliberately minimized in order to focus on energetic aspects of harmonicity-based unmasking. 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 reference without any harmonicity or amplitude modulation. The monotonized and intonated buzzes are non-speech maskers without

amplitude modulation but with harmonicity (F0 fixed or varying over time). Vocoded speech has amplitude modulation 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. Reversed speech was added as a control to confirm that our forward (natural) speech masker had minimal IM: since reversed speech has very little IM then we expected no difference between these two maskers.

2.2 Methods

Listeners

Nine listeners (ages 19-23 years, mean age 21) participated in this 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 subjects had normal pure tone thresholds 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.

Stimuli

The target sentences used 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). 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 spectrum as the monologue. The buzzes were harmonic complexes with partials in random phase, that were passed through a finite impulse response filter to match the average long-term excitation pattern (Glasberg and Moore, 1990) of the monologue. 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 and Weenink, 2018). The monotonized speech was created by fixing the F0 to the mean F0 (112 Hz) of the monologue using PRAAT PSOLA. The vocoded speech was created with an 8-channel vocoder, using an envelope low-pass filter-cutoff frequency of 150 Hz and also had the same average long-term excitation pattern as the speech. Figure IV.1 presents the excitation patterns of all maskers.

Stimuli were spatialized using anechoic KEMAR head related transfer-functions (HRTFs; Gardner and Martin, 1994). 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 the talker sex, and the difference in the structure and content of the speech materials, all served to distinguish the target from the maskers and thus minimize IM.

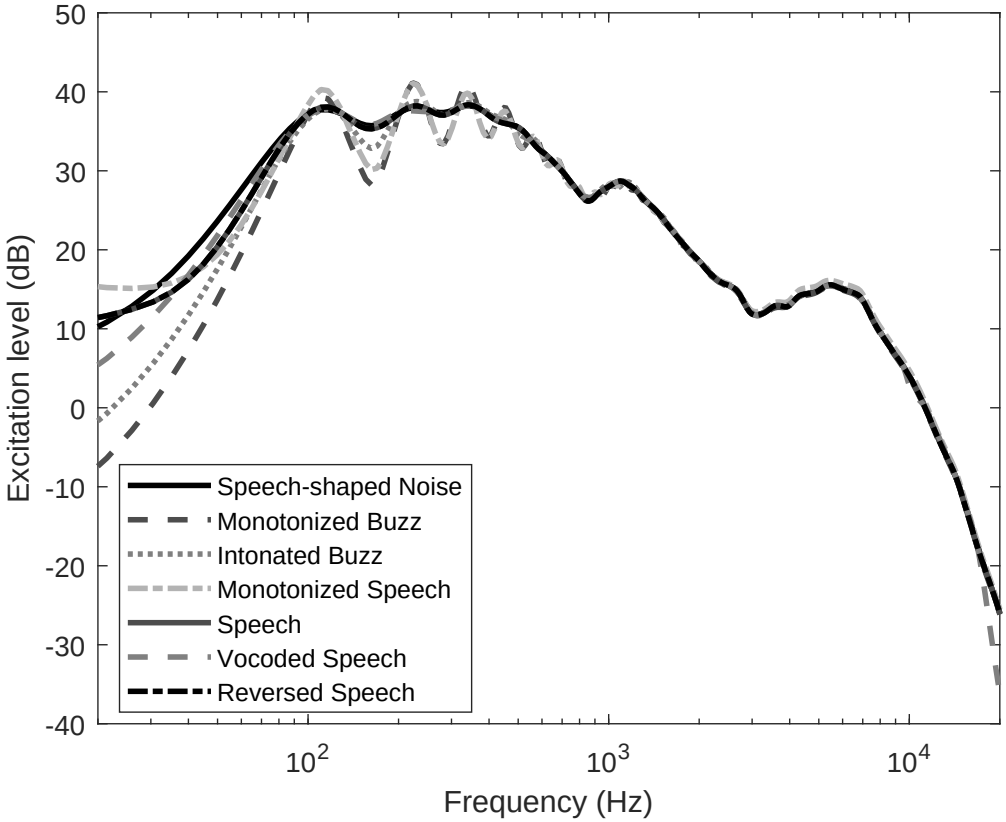

Fig. IV.1 Long-term excitation patterns of the seven maskers tested in the main experiment

Procedure

The stimuli were presented with a 24-bit soundcard (RME HDSP 9632, Haimhausen, Germany) via 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). The participants 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 SNR varied randomly between five chosen values (from -40 to -10 dB). The masker level was fixed at 65 dB SPL. At the beginning of each session, there was always a block 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. Logistic functions were fitted using the psignfit toolbox version 4 for MATLAB, which implements the maximum likelihood method described by Wichmann and Hill (2001). The lower asymptote was set to chance performance (12.5 %). SRTs (SNRs corresponding to 50% words correct) were extracted from the logistic fits.

2.3 Results

All subjects (except the one who was excluded) performed well in quiet with scores about 85% correct at all target levels. Figure IV.2, panel A, 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). SRTs were lower for separated than for co-located conditions for all masker types.

Tukey pairwise comparison indicated that in the co-located condition, SSN led to a significantly higher SRT than all other masker types. Both intonated and monotonized buzzes led to significantly higher SRTs than amplitude-modulated maskers. The monotonized buzz was significantly lower than the intonated buzz. There was no significant difference in SRT between monotonized speech, speech, vocoded speech and reversed speech.

In the separated condition, the difference in SRT between SSN and intonated buzz was not significant. SSN and intonated buzz led to significantly higher SRTs than the other masker types. Monotonized buzz led to SRTs significantly lower than SSN and intonated buzz and significantly higher than amplitude-modulated maskers. There was once again no significant difference in SRT among the reversed, vocoded, monotonized and natural speech maskers.

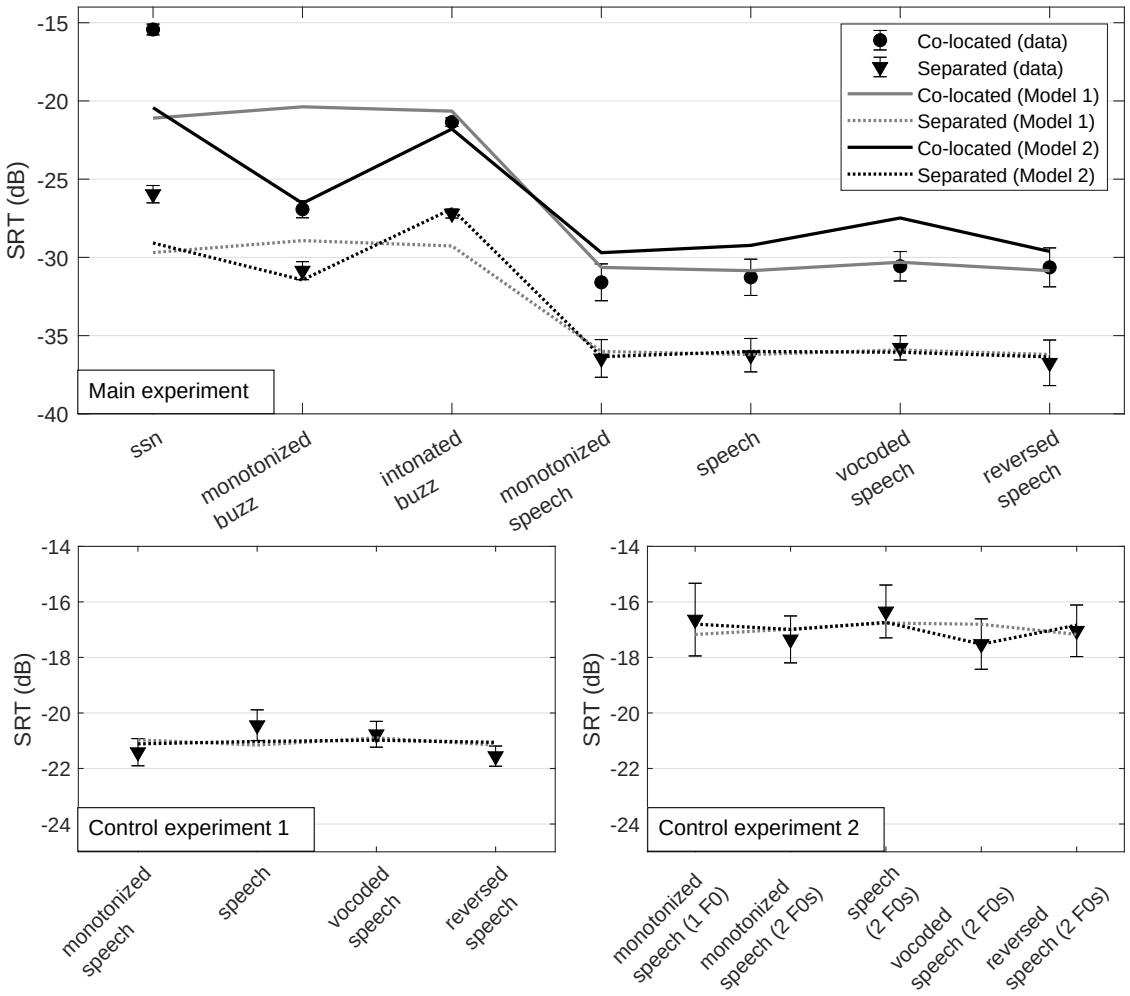

Fig. IV.2 Mean SRTs with standard errors across participants measured in the main experiment (top panel), in the control experiment 1 (bottom left panel), and control experiment 2 (bottom right panel) with the corresponding model predictions using: model 1, a binaural model without harmonic cancellation (Vicente and Lavandier, 2020) and model 2, a binaural model with harmonic cancellation.

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 the 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, this was achieved by making the speech task more difficult by using open set target sentences as opposed to a closed set matrix. In control experiment 2, in addition to using open-set materials, we also added a second speech masker to increase EM.

The idea behind the control experiments was that it could have been possible that there was not enough masking to begin with to observe an effect of harmonicity in the main experiment. If there were already a large masking release thanks to the release cues mentioned above, there would be no need for another mechanism like harmonic cancellation. The aim of control experiment 2 was to add more EM so there would be more opportunities for harmonic cancellation. This was done by replicating some conditions of the main experiment, with two maskers instead of one, so that EM increases. By adding another masker, there would be more voiced parts overall in the speech masker and less amplitude modulation, which could increase opportunities for harmonic cancellation and at the same time give less opportunities for temporal glimpsing. We designed this experiment so that IM, which is known to be more important for 2-voice than 1-voice maskers (Brungart et al., 2001; Freyman et al., 2004), was still minimized as much as possible.

3.2 Methods

Five of the listeners (ages 19-22 years, mean age 21) who participated in the main experiment also participated in control experiment 1. Five different listeners (ages 19-22 years, mean age 21; normal pure tone thresholds at octave frequencies from 250 to 8000 Hz) participated in control experiment 2.

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. Once again, target and maskers were different speech materials, with different accents, mean F0 and were always spatially separated (target at 0° and masker at 60°) to minimize IM. In control experiment 2, the two maskers were simulated at the same location.

In control experiment 1, only four masker conditions were re-tested (because they were the conditions that could have been influenced by a floor effect): speech, monotonized speech, vocoded speech, 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-voices natural speech, two-voices monotonized speech, two-voices vocoded speech, two-voices reversed speech. An additional fifth masker type, a single-voice monotonized speech, was added. It was constructed by adding two parts of the monotonized male monologue from the main experiment. This resulted in a two-talker masker with a steady F0 at 112 Hz. This single-voice monotonized speech was the masker that had the more chances to be canceled by harmonic cancellation in theory because it has more voiced parts than the monotonized speech from the main experiment, less temporal dips, and has a single steady F0. Reversed speech was kept here as a control for IM.

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 in each session.

3.3 Results

Figure IV.2 presents the SRTs measured in control experiments 1 and 2, in the bottom left and right panels, respectively. The thresholds in the control experiments were higher than in the main experiment as intended. In both control experiments, the ANOVA did not reveal any significant effect of masker type across the reversed, vocoded, monotonized and natural speech maskers.

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 and Lavandier (2020) to take into account modulated

noise maskers, and the binaural harmonic cancellation model described in chapter III. By comparing how well each model can account for the data, we were able to provide further support for (or against) a role for harmonic cancellation in reducing EM for these stimuli.

4.2 Models

Model 1 computes the better-ear SNR and a binaural unmasking advantage in each frequency band. It operates over short time frames (24 ms for the better ear and 300 ms for the binaural unmasking). This model is able to accurately predict binaural differences and dip listening for noise maskers (Vicente and Lavandier, 2020). Model 2 corresponds to model 4 presented in chapter III. It operates over time frames of 300 ms. In each time frame, if the masker signal has an F0, harmonic cancellation is applied (comb filter tuned to the mean F0 of the masker in that time frame) and only the better-ear SNR is computed. If there is no F0, the model computes the better-ear SNR and binaural unmasking advantage. In this model, harmonic cancellation and binaural unmasking are assumed to be mutually exclusive. The condition for applying harmonic cancellation was that the masker signal was voiced at least 50 % of the time in the time frame. This model proved useful to predict the effects of SRM, intonation and amplitude modulations in the masker for buzz maskers (see chapter III).

The models were applied to the three experiments. The reference chosen to fit the predictions to the data were the average SRT across conditions for each experiment.

4.3 Results

Figure IV.2 (top panel, gray lines) presents the model predictions for the main experiment using model 1. This model did not predict the differences between SSN, monotonized buzzes and intonated buzzes. It predicted a difference of about 10 dB between SSN and amplitude-modulated maskers. This predicted dip listening is underestimated compared to the effect observed in the data. The model accurately predicted the SRM for SSN and amplitude-modulated maskers, but it overestimated the SRM for buzz maskers.

Figure IV.2 (top panel, black lines) presents the model predictions using model 2. This model accurately predicted the differences between intonated and monotonized buzzes. It accurately predicted the differences between buzzes and amplitude-modulated maskers. However, it underestimated the differences between SSN and the other masker types. The predicted SRM was similar to that observed in the data, although it was slightly overestimated for amplitude-modulated maskers.

Figure IV.2 (bottom panels) presents the model results for the control experiments using model 1 (gray lines) and model 2 (black lines). Model 1 did not predict differences between

masker types for either of the control experiments. Model 2 did not predict any differences between masker types for control experiment 1. It predicted a slightly lower SRT for vocoded speech in control experiment 2 (less than 1 dB difference).

5 Discussion

5.1 Harmonicity

In the behavioral experiment, SRTs were highest for the SSN masker, which had neither harmonicity nor amplitude modulation. Compared to buzzes, which also had no amplitude modulation, the 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 and Rosen (2015), who measured SRTs for speech against SSN and intonated harmonic complexes and found that SRTs were about 9 dB 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. Model 1 did not predict the difference between monotonized buzzes, intonated buzzes and SSN. Contrary to model 1, model 2 accurately predicts the differences between monotonized buzzes and intonated buzzes. This is in agreement with the modeling results of chapter III, which suggests that spectral glimpsing alone cannot explain this difference and that harmonic cancellation is needed in order to predict the effect. However, model 2 underestimated the difference between SSN and buzzes by 4.6 dB. The only previous study that tried to predict the difference between noise and intonated buzz also underestimated this effect by about 5 dB (Steinmetzger et al., 2019).

Our results are in agreement with Leclère et al. (2017). With natural speech as a target, they found a difference in SRT of about 4-5 dB between intonated and monotonized buzzes. In our data the difference is about 5-6 dB. Other studies also found that intonated buzzes led to higher SRTs than monotonized buzzes (Deroche and Culling, 2011; Green and Rosen, 2013). As proposed by Leclère et al. (2017), the difference between intonated and monotonized maskers could be due to limitations in a harmonic cancellation mechanism. Specifically, if this mechanism is at play for these stimuli, it may be that a masker with a steady F0 is easier to cancel or that it is more difficult to "follow" the F0 contour when it varies over time. 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 provides enough masking release so that there is no further benefit to be gained from the weak harmonicity cue in the intonated condition.

We did not find a significant difference between speech and monotonized speech in the main experiment, nor in the control 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. This could be explained by the fact that speech contains unvoiced segments and amplitude modulations, which results in fewer opportunities for harmonic cancellation than with continuous harmonic complexes. 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. An alternative explanation might be that when there is already a release from masking due to another mechanism (amplitude modulation in this case), the advantage due to harmonicity is reduced or even negligible. One interesting result is that the harmonic-cancellation model 2 does not predict an advantage for monotonized speech compared to naturally intonated speech, as opposed to the advantage it predicts for monontonized buzzes over intonated buzzes. This seems to corroborate the hypothesis that, with speech maskers, the presence of unvoiced parts and amplitude modulations significantly reduces the influence of harmonic cancellation. The model predicted a small difference between vocoded speech, which was unvoiced, and the three other speech-like maskers that was not observed in the data.

One concern we had 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 may 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 maskers 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 monotonized speech was not significantly different from the SRT 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. The model predictions once again seem to support that hypothesis.

Note that although we found no significant differences between SRTs for the different amplitude-modulated maskers, we cannot provide definitive evidence for the lack of an effect of masker type. Bayesian statistics would be required to prove that the effect does not exist

(Keysers et al., 2020), but such statistics were not conclusive when used here, possibly due to the limited number of listeners involved. However, we can say that if there is any effect, it is most probably small.

Our conclusion is that harmonicity-based effects on EM are negligible for speech-on-speech situations, and that examples of harmonicity-based release from masking reported in the literature for speech maskers (Brokx and Nooteboom, 1982; Deroche and Gracco, 2019; Popham et al., 2018) likely reflect a release from IM, as F0 is a strong cue for speech stream segregation (David et al., 2017). 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., 2020). The modeling results presented here seem to confirm this and suggest that when not considering harmonic complexes as maskers, the "modulated-noise" model from Vicente and Lavandier (2020) is likely to be sufficient.

5.2 Spatial separation

The results showed a main effect of spatial separation between target and masker, consistent with a large body of literature (Beutelmann and Brand, 2006; Collin and Lavandier, 2013; Hawley et al., 2004; Leclère et al., 2017). SRM was observed for all masker types, but it differed in magnitude. The SRM was larger for SSN (10 dB) than for all of the other masker types. This may be because there was more masking to begin with for SSN, and thus more potential for masking release. For buzzes, the SRM was smaller for the monotonized buzz (4 dB) than the intonated buzz (5.8 dB). This is consistent with the results from Leclère et al. (2017), who found a difference in SRM of 0.8 dB between intonated and monotonized buzzes. One potential explanation could be that there was already release from masking (e.g., due to harmonic cancellation for monotonized buzzes) which means that there was less potential for SRM. Related to this, Leclère et al. (2017) suggested that this difference might have been due to a floor effect, which limited the SRM for monotonized buzzes, as the SRTs were already very low as a result of the F0 difference.

The SRM predicted by model 1 was equivalent for SSN and buzzes, because this model cannot distinguish buzzes from noises apart from their minor differences in long-term spectrum (figure IV.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 was in good agreement with the observed SRM. This reduction in SRM for amplitude-modulated maskers is due to the ceiling set at 20 dB in the model. As the dip listening already gave a large advantage, the predicted SRTs could not go lower.

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 (see chapter III). Without this assumption, the model would overpredict SRM for harmonic maskers. Our rationale for 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 these two mechanisms 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. Collin and Lavandier (2013) found a difference between SRTs measured for noise and single-talker-modulated noise of about 5 dB. Beutelmann et al. (2010) found a difference of 11.2 dB, which is closer to our results. The target speech material might explain those differences. Beutelmann et al. (2010) also used closed set matrix sentences (five words, ten choices), whereas Collin and Lavandier (2013) used open-set sentences. There is some evidence that the benefit derived from amplitude fluctuations in the masker depends on the nature of the target speech materials (Best et al., 2019; Schoof and Rosen, 2015). Model 1 did not fully predict the observed SRT difference between SSN and the four amplitude-modulated maskers. A very similar model was able to accurately predict the SRT difference measured between SSN and vocoded speech in co-located and separated conditions using open-set target sentences (Vicente et al., 2020). The difference in prediction accuracy might be explained by the fact that the dip listening advantage observed here was larger because of task and the speech material used. The models considered here could not predict such an effect of the speech material or task difficulty.

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 and Rosen (2015) also found that the advantage due to amplitude modulation was smaller than the advantage due to periodicity. Our results provide a slightly different picture, in that we found evidence for temporal dip listening but no strong evidence for release due to harmonicity for speech maskers. 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 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. The very similar SRTs obtained for speech and reversed-speech maskers indicate that IM was minimized in this study. SRTs measured for unmodulated sounds (SSN and buzzes) suggested an advantage due to harmonicity that is impaired by intonation (in agreement with Leclère et al., 2017). Such conditions continue to provide the most compelling case for a harmonic cancellation mechanism. On the other hand, the results for amplitude-modulated maskers suggest a very limited role for harmonic cancellation for combating EM in mixtures of talkers. This conclusion was further supported by the predictions of models with and without a harmonic cancellation component.

General conclusions

This PhD project aimed towards developing a speech intelligibility model that could account for energetic effects of speech masked by speech, in particular the contribution of F0-based effects. Of particular interest was the contribution of harmonic cancellation in speech-on-speech situations.

A model was proposed to predict speech intelligiblity against harmonic maskers. The mechanisms of spectral glimpsing and harmonic cancellation were both included in the model. It was used to describe the data of two experiments (Deroche et al., 2014b) in which speech reception thresholds were measured for stationary harmonic maskers varying in their F0 and degree of harmonicity. Key model parameters (jitter in the masker F0, shape of the cancellation filter, frequency limit for cancellation, and signal-to-noise ratio ceiling) were optimized by maximizing the correspondence between the predictions and data. The model was able to accurately describe the effects associated with varying the masker F0 and harmonicity.

The model was further developed so that it could take into account F0 variations over time, binaural differences, and amplitude modulations in the masker. This was done by segmenting the masker signal into time frames, applying the harmonic cancellation within each time frame, before computing the better-ear SNR and binaural unmasking advantage, as done in previous models from the literature. This new model version was tested on two experiments involving harmonic complexes maskers that varied in spatial location, temporal envelope and F0 contour (Leclère et al., 2017). Interactions between unmasking effects could be taken into account in the model by varying the time frame duration and excluding the binaural unmasking computation when harmonic cancellation was active. The model was able to accurately predict the effects associated with spatial separation, intonation, amplitude modulations and F0 differences. The modeling results showed that the implementation of

harmonic cancellation was necessary to predict F0-based effects and their interaction with spatial unmasking and dip listening.

A behavioral experiment was conducted to investigate the potential role of harmonic cancellation in masked speech intelligibility. While there is evidence that harmonic cancellation plays a role against harmonic complex maskers, its utility against natural speech maskers with non-stationary F0s and unvoiced parts was investigated. Speech reception thresholds were measured using seven types of masker: speech-shaped noise, monotonized and intonated harmonic complexes, monotonized speech, noise-vocoded speech, reversed speech and natural speech. Those stimuli provided comparison between conditions with and without harmonic structure, amplitude modulation and variations in F0 over time. This study confirmed the results from the literature that, for harmonic complexes, the advantage due to harmonicity is impaired by intonation in the masker F0. This study also suggested that the contribution of harmonic cancellation in situations of speech masked by speech is probably very limited. This conclusion was supported by the predictions of the harmonic cancellation model.

This PhD work allowed to develop a speech intelligibility model for harmonic complex maskers. It also provided further knowledge on the role of harmonic cancellation in cocktail party situations, with speech maskers, suggesting that F0-based effects in such situations could be mainly due to a release from IM. The work also raised some questions on the potential interaction between F0-based effects and binaural unmasking. Additional studies would be needed to better understand this. Overall, this work suggests that speech intelligibility models validated for modulated noise maskers should be sufficient to give a first approximation of predictions for speech maskers when only EM effects are involved.

Résumé en français

1 Introdution

Dans la vie de tous les jours, il existe de nombreuses situations au cours desquelles nous sommes confrontées au défi d'écouter un discours en présences de sources concurrentes: dans les transports en commun, les restaurants, les salles de classe, les bureaux en open space... Plusieurs mécanismes qui permettent au système auditif de améliorer l'intelligibilité de la parole dans de telles conditions ont été identifiés dans la littérature. De tels mécanismes reposent sur les propriétés acoustiques de la source cible et des sources masquantes. De nombreuses études se sont intéressées à des stimuli et des situations moins complexes que celles que l'on peut rencontrer dans la vie réelle (plusieurs locuteurs et bruits concurrents) afin de simplifier le problème. A l'heure actuelle, il reste encore à déterminer dans quelle mesure ces résultats peuvent s'appliquent à des situations réelles. Afin d'améliorer l'intelligibilité

de la parole dans de telles situations, il est important de mieux comprendre les mécanismes perceptifs impliqués. Les modèles d'intelligibilité de la parole constituent un moyen d'y parvenir. Une modélisation précise des mécanismes utilisés par le système auditif pour améliorer l'intelligibilité de la parole peut conduire à une meilleure compréhension de ces mécanismes. Ces modèles peuvent ensuite être utilisés pour améliorer la conception des bâtiments ou la conception d'aides auditives par exemple. Bien que plusieurs modèles aient été proposés dans la littérature, aucun d'entre eux n'est pour l'instant capable de prédire l'intelligibilité de la parole dans des situations complexes et plus réalistes, en présence de plusieurs sources concurrentes de parole. Afin de créer un tel modèle, plusieurs caractéristiques de la parole doivent être prises en compte. En particulier, les effets associés aux fréquences fondamentales (F0s, la propriété acoustique associée à la hauteur tonale) des signaux de la cible et du masqueur nécessitent d'être étudiés plus spécifiquement.

Cette thèse est composée de quatre chapitres : un état de l'art des connaissances sur la problématique de l'intelligibilité de la parole en présence de sources concurrentes de parole et la modélisation des effets associés à la F0. Les chapitres suivants présentent le développement d'un modèle d'intelligibilité de la parole avec l'implémentation de la suppression harmonique pour des complexes harmoniques puis son extension à des complexes harmoniques dont la F0, l'enveloppe temporelle ou la localisation varient. Une troisième étude présente une expérience menée dans le but d'examiner le rôle de la suppression harmonique sur différents types de masqueurs allant du bruit à la parole. L'application du modèle développé dans les chapitres précédents a permis de confirmer les conclusions expérimentales concernant le rôle de la suppression harmonique pour l'intelligibilité de la parole dans les situations de cocktail party.

2 Etat de l'art

Dans un environnement bruyant, il existe plusieurs facteurs pouvant empêcher les auditeurs de bien comprendre la personne qui parle, comme les bruits parasites et les locuteurs concurrents. Dans ce genre de situation, souvent appelé le "cocktail party problem" citep Cherry1953, les auditeurs sont confrontés à de nombreux défis. Ils doivent d'abord séparer les différentes sources, puis sélectionner la source sur laquelle se concentrer et enfin comprendre les informations véhiculées par cette source sonore. L'un des facteurs qui peuvent affecter l'intelligibilité de la parole est le rapport signal/bruit (signal to noise ratio en anglais, SNR) auquel les sons sont présentés. Une façon d'étudier l'intelligibilité de la parole consiste à mener des expériences dans lesquelles un signal de parole est présenté aux auditeurs en présence d'une source masquante et les auditeurs doivent rapporter la phrase qu'ils ont entendue. Une mesure d'intelligibilité de la parole peut être obtenue en comptant le pourcentage de mots que les auditeurs ont compris. Le seuil de réception de la parole (speech reception threshold en aglais, SRT) correspond au SNR auquel un auditeur est capable de comprendre une certaine proportion des mots cibles (souvent fixé à 50 %). Une diminution de SRT correspondra donc à une amélioration de l'intelligibilité.

Le cas des cocktail party est particulièrement complexe à étudier car les sources masquantes ne sont pas seulement du bruit mais aussi des sources de parole, et les propriétés acoustiques sont différentes dans la parole et le bruit. Certaines parties du signal de parole sont harmoniques, ce qui signifie que leur spectre est composé d'harmoniques qui sont espacées en fréquence à des multiples de la fréquence fondamentale (F0). Il est composé de voyelles (qui sont des complexes harmoniques) et de consonnes. Certaines des consonnes sont voisées (harmoniques) tandis que d'autres non voisées (pas harmoniques). La parole présente égale-

ment des modulations d'intonation (variation de la F0 au cours du temps) et d'amplitude. Ainsi, dans le cas particulier de sources concurrentes de parole, de nombreux facteurs qui ne sont pas présents avec le bruit entrent en jeu.

Le masquage est un élément majeur pour comprendre le problème des cocktails. Il décrit la manière dont les sources concurrentes empêchent les auditeurs de comprendre un discours, et il est souvent quantifié en termes de SRT. Le masquage peut être séparé en deux catégories: le masquage énergétique et le masquage informationnel citep Brungart2001a. Le masquage énergétique (EM) fait référence au masquage qui se produit lorsque les signaux cible et masquant se chevauchent et entrent en compétition à la périphérie du système auditif citep Durlach2003, c'est-à-dire lorsque les signaux cible et masqueur se chevauchent dans les domaines temporel et fréquentiel. Dans les situations de cocktail et dans toute situation où un auditeur essaie de comprendre un locuteur parmi des locuteurs concurrents, le masquage peut se produire même si la cible est suffisamment audible. Le masquage informationnel (IM) fait référence à une réduction de l'intelligibilité de la parole qui ne peut pas être expliquée par l'EM. Plusieurs facteurs qui empêchent l'auditeur de séparer le discours cible des voix concurrentes, ou de se concentrer sur le bon locuteur.

Le système auditif est capable d'utiliser plusieurs mécanismes pour réduire le masquage et améliorer l'intelligibilité de la parole en présence de sources sonores concurrentes. Ces mécanismes reposent sur des signaux acoustiques tels que F0 et la structure harmonique, les différences binaurales et les fluctuations d'enveloppe temporelle.

Plusieurs études ont montré une amélioration de l'intelligibilité de la parole lorsque la source cible et une source masquante harmonique ont des F0 différentes. Deux mécanismes ont été proposés pour expliquer cela. Le système auditif pourrait être capable de capter de l'information dans les trous spectraux du masqueur harmonique (mécanisme de "spectral glimpsing"). Le mécanisme de suppression harmonique suggère que le système auditif serait capable de repérer la structure harmonique du masqueur et la supprimer lorsque sa F0 est différente de celle de la source cible.

Lorsque le signal masquant présente des modulations d'amplitude les auditeurs pourraient être capables de capter de l'information dans les trous de modulation (mécanisme d'écoute dans les trous) afin d'améliorer l'intelligibilité.

Lorsque les sources cibles et masquantes sont situées à des endroits différents, l'intelligibilité est également améliorée grâce aux différences interaurales de niveau (interaural level differences en anglais, ILDs) et aux différences interaurales de temps (interaural time differences en aglais, ITDs). Deux mécanismes permettent d'exploiter ces différences interaurales et donc d'améliorer l'intelligibilité : l'écoute à la meilleure oreille et le démasquage binaural.

3 Modèle de prédiction de l'intelligibilité en présence d'un masqueur harmonique

Alors que plusieurs modèles sont disponibles pour prédire l'intelligibilité de la parole pour divers types de bruit masquants, il n'y a actuellement aucun modèle permetant de prédire les effets de l'harmonicité et des différences de F0.

Le but de la cette étude était de développer un modèle d'intelligibilité capable de prédire les effets ΔF0 en présence d'un masqueur harmonique en incluant le mécanisme de suppression harmonique dans un modèle existant(basé sur Collin and Lavandier, 2013).

L'implementation de la suppression harmonique dans le modèle est faite en plusieurs étapes. La F0 du signal masquant est estimée, puis un filtre en peigne (supprimant la F0 et ses harmoniques) est créé. Ce filtre est appliqué à la fois au signal cible et au signal masquant. Si cela donne un avantage lors du calcul du SNR par bande de fréquence, on applique la suppression harmonique. Quatre paramètres ont été introduits dans le modèle : une variation aléatoire dans l'estimation de la F0, un paramètre contrôlant la forme du filtre, une limite en fréquence jusqu'à laquelle la suppression harmonique est appliquée, et un plafond pour limiter le SNR. Une analyse des quatre paramètres introduits dans le modèle a été effectuée et des valeurs ont été choisies pour optimiser les performances du modèle. Cette analyse a été effectuée en utilisant les données de l'expérience tirée de Deroche et al. (2014b) qui utilise des masqueurs harmoniques complexes stationnaires avec une F0 fixe et aucune modulation d'amplitude.

Le modèle décrit avec une très bonne précision les SRTs mesurés dans deux expériences tirées de Deroche et al. (2014b) utilisant des masqueurs variant en F0 et en degré d'harmonicité. L'erreur moyenne et l'erreur maximum entre les données de l'expérience et les prédictions étaient inférieures à 1 dB, et similaires à celles rapportées précédemment pour d'autres modèles d'intelligibilité de la parole. Le modèle ne capture pas complètement certains des effets observés dans les données comme l'interaction entre la F0 du masqueur et l'harmonicité.

4 Extension du modèle pour un masqueur variant en intonation, enveloppe temporelle ou localisation

Le but de cette étude est d'étendre le modèle d'intelligibilité de la parole proposé par Prud'homme et al. (2020) à des masqueurs plus complexes présentant de l'intonation, les

indices binauraux et des modulations d'amplitude. L'approche adoptée a été d'essayer différentes versions de modèles pour étudier les interactions entre les effets (effets basés sur la F0, la localisation, écoute dans les trous) avec des masqueurs plus complexes. Cette approche a permis d'estimer l'influence des différents mécanismes sur le masquage énergétique de la parole. Dans ce chapitre, la modélisation n'est pas seulement utilisée à des fins de prédiction, mais aussi pour confirmer les hypothèses suggérées par les résultats expérimentaux.

Différentes versions de modèles — variations des modèles de Vicente and Lavandier (2020) et Prud'homme et al. (2020) — ont été testées sur les stimuli de Leclère et al. (2017), qui sont des complexes harmoniques variant dans leur F0, localisation spatiale ou modulation d'amplitude. Ces masques sont plus complexes que ceux testés par Prud'homme et al. (2020) (qui étaient des complexes harmoniques stationnaires et monotones de Deroche et al., 2014b), tout en étant plus simples que des signaux de parole. Ainsi, ils représentent une étape entre les complexes harmoniques stationnaires et la parole, permettant d'étudier comment les différentes versions du modèle prennent en compte l'intonation, les modulation d'amplitude et les différences binaurales, tout en évitant pour l'instant les autres problèmes complexes de parties non voisées et de masquage informationnel.

Quatre modèles ont été testés :

- Model 1: modèle non-stationnaire, binaural, sans suppression harmonique (Vicente and Lavandier, 2020).
- Model 2: modèle diotique avec suppression harmonique (Prud'homme et al., 2020).
- Model 3: modèle non-stationnaire, binaural, avec suppression harmonique.
- Model 4: modèle non-stationnaire, binaural, avec suppression harmonique dans lequel le démasquage binaural et la suppression harmonique sont mutuellement exclusifs. Le démasquage binaural n'est pas calculé lorsque la suppression harmonique est appliquée.

Le modèle de Prud'homme et al. (2020) a été modifié afin de prendre en compte les variations de F0 au cours du temps et les différences binaurales (modèles 3 et 4). Un nouveau paramètre (durée de la fenêtre temporelle) a eu le plus d'influence pour l'expérience 2 de Leclère et al. (2017) qui impliquait des masqueurs modulés en amplitude, ainsi différentes tailles de fenêtre temporelle ont été testées pour cette expérience. Cela nous a permis de définir une intervalle de valeurs acceptables pour ce paramètre qui a donné de bonnes prédictions pour l'expérience 2. Certaines de ces valeurs ont ensuite été testées pour l'expérience 1 de Leclère et al. (2017), qui impliquait des conditions binaurales pour trouver la valeur finale

qui correspondait le mieux aux données expérimentales pour les deux expériences. La comparaison des prédictions des différents modèles a montré que la suppression harmonique était nécessaire afin de prédire les effets associés à la F0 (différences d'intonation et différences de F0). Les modèles binauraux prennant en compte le démasquage binaural (modèles 1 et 3) surestimaient le démasquage spatial observé dans les données. En revanche le modèle 4, qui prend en compte uniquement l'écoute à la meilleure oreille, prédit correctement le démasquage spatial.

Le modèle proposé ici donne des prédictions précises pour les deux expériences (erreur moyenne < 0.6 dB et erreur maximale < 1.5 dB) avec des masqueurs harmoniques ayant des F0 différentes, des contours de F0 différents, des modulations d'amplitude et une séparation spatiale. Le mécanisme de suppression harmonique était nécessaire afin de prédire les effets basés sur F0 et les prédictions étaient les meilleures lors de l'utilisation du modèle avec une taille de fenêtre de 300 ms. Cependant, la version la plus réussie du modèle repose sur l'hypothèse que les mécanismes suppression harmonique et de démasquage binaural sont mutuellement exclusifs, ce qui soulève des questions sur l'additivité de ces effets et pourraient donner lieu à une étude plus approfondie.

5 Le rôle de la suppression harmonique dans l'intelligibilité de la parole en présence de sources concurrentes de parole

Afin de mieux comprendre les effets de F0 sur l'intelligibilité de la parole, plusieurs études ont simplifié le problème en utilisant des complexes harmoniques — souvent appelés buzz — au lieu de la parole comme masqueurs (Deroche and Culling, 2013; Leclère et al., 2017; Steinmetzger and Rosen, 2015). Ces études confirment que, même lorsque les masqueurs sont des stimuli simple caucant principalement de l'EM, l'harmonicité joue un rôle dans la réduction du masquage. Il n'est pas clair cependant si ces conclusions peuvent s'appliquer aux cas plus complexes des masqueurs de parole. En effer, certaines caractéristiques du signal qui ne sont pas présentes dans les complexes harmoniques sont à prendre en compte avec la parole: la présence de segments non voisés, l'intonation et la modulation d'amplitude. Certaines études indiquent que ces facteurs pourraient réduire le rôle de l'harmonicité pour les situations impliquant des sources concurrentes de parole (Deroche and Culling, 2013; Deroche and Gracco, 2019; Leclère et al., 2017).

L'objectif de la cette étude était d'étudier si l'harmonicité (et la suppression harmonique en particulier) joue un rôle dans la réduction de l'EM dans les situations de sources concur-

rentes de parole. La plupart des études précedemment réalisées dans la littérature ont testé les conditions avec des masqueurs de parole, buzz ou bruit séparément. Les but était ici de tester ces types de masqueurs dans les mêmes conditions et sur les mêmes sujets.

L'intelligibilité de la parole a été mesurée en présence de différents types de masqueurs, allant du bruit à la parole. En comparant les SRT entre les types de masqueurs, l'objectif était d'estimer l'influence de l'harmonicité tout en contrôlant les diverses caractéristiques de la parole. Dans cette expérience, l'IM a été minimisée afin de se concentrer uniquement sur les aspects énergétiques du démasquage dû à la F0. Sept types de masqueurs ont été utilisés : un bruit ayant le même spectre que la parole, un buzz monotonisé, un buzz intonisé, une voix monotonisée, une voix naturelle, une voix vocodée et une voix inversée. Le bruit a été utilisé comme référence sans structure harmonique ni modulation d'amplitude. Les buzz monotonisés et intonisés sont des masqueurs sans modulation d'amplitude mais avec une structure harmonique (F0 fixe ou variable dans le temps). La parole vocodée a une modulation d'amplitude similaire à la parole, ce qui devrait mener à un avantage dû à l'écoute dans les trous, mais aucune harmonicité. La parole monotonisée a été incluse car des données de la littérature suggérèrent que la suppression harmonique fonctionnait plus efficacement pour les masqueurs harmoniques monotonisés que les masqueurs harmoniques intonisés. La parole inversée a été ajoutée comme contrôle pour confirmer que le masqueur de parole (naturel) produisait un IM minimal : puisque la parole inversée a très peu d'IM, ces deux masqueurs ne devraient pas présenter de différences dans les résultats. Deux conditions ont été testées pour chaque masqueur : co-localisé (la cible et le masqueur sont présentés au même endroit, face à l'auditeur) ou séparé (la cible est présentée face à l'auditeur et le masqueur à 60° sur le côté).

Les différentes types de masqueurs ont permis de comparer des masqueurs avec et sans structure harmonique, modulation d'amplitude et variations de F0 dans le temps. Les SRTs obtenus pour les masqueurs de parole et de parole inversée indiquent que l'IM a été minimisée dans cette étude. Les SRTs mesurés pour les sons non modulés (SSN et buzz) suggèrent un avantage dû à l'harmonicité qui est néanmoins altérée par l'intonation (en accord avec les résultats de Leclère et al., 2017). Ces résultats apportent donc des éléments suplémentaires suggérant un rôle de la suppression harmonique pour ces conditions. D'autre part, les résultats des masqueurs modulés en amplitude suggèrent un rôle très limité de la suppression harmonique pour l'EM en présence de sources concurrentes de parole puisque le effets observés sur les buzz ne sont pas observés avec la parole naturelle et monotonisée et qu'il n'y avait pas de différence significative entre la parole et le vocodeur. Deux expériences contrôles, réalisées afin de rendre la tâche plus complexe et d'augmenter l'EM (afin d'augmenter la

possibilité d'utiliser la suppression harmonique pour les auditeurs) ont été réalisées et ont confirmé les résultats de l'expérience principale.

Le modèle proposé dans le chapitre précedent, ainsi que le modèle proposé par Vicente and Lavandier (2020) ont été appliqués aux stimuli de l'expérience afin de confirmer les conclusions expérimentales sur le rôle de la suppression harmonique. Les prédictions des modèles suggèrent un rôle de la suppression harmonique pour les buzz mais pas pour les masqueurs de parole. Pour une première approximation, les modèles créés pour prédire l'intelligibilité en présence de bruit uniquement peuvent donc être utilisés pour des masqueurs de parole.

6 Conclusions

Ce projet visait à développer un modèle d'intelligibilité de la parole qui pourrait prendre en compte des effets énergétiques pour les sources concurrentes de parole, en particulier la contribution des effets en lien avec la F0 (notamment la suppression harmonique).

Un modèle a été proposé pour prédire l'intelligiblité de la parole en présence de masqueurs harmoniques. Les mécanismes de spectral glimsping et de suppression harmonique ont tous deux été inclus dans le modèle. Il a été utilisé pour décrire les données de deux expériences (Deroche et al., 2014b) dans lesquelles des SRTs ont été mesurés pour des masqueurs harmoniques stationnaires variant dans leur F0 et leur degré d'harmonicité. Les paramètres clés du modèle (variation dans la F0 estimée du masqueur, forme du filtre, limite en fréquence et plafond pour le SNR) ont été optimisés en maximisant la correspondance entre les prédictions et les données. Le modèle proposé décrit avec précision les effets associés aux variations de la F0 et l'harmonicité du masqueur.

Le modèle a été amélioré de manière à pouvoir prendre en compte les variations de F0 dans le temps, les différences binaurales et les modulations d'amplitude dans le masqueur. Cela a été fait en segmentant le signal masquant en fenêtres temporelles, en appliquant la suppression harmonique dans chaque fenête temporelle, avant de calculer le SNR de la meilleure oreille et l'avantage de démasquage binaural, comme cela a été fait dans les modèles précédents de la littérature. Cette nouvelle version du modèle a été testée sur deux expériences impliquant des masqueurs de complexes harmoniques qui variaient dans leur localisation, leur enveloppe temporelle et leur contour de F0 (Leclère et al., 2017). Les interactions entre les effets de démasquage sont prises en compte par le modèle en faisant varier la durée de la fenêtre temporelle et en excluant le calcul de démasquage binaural lorsque la suppression harmonique était active. Le modèle a pu prédire avec précision les

effets associés à la séparation spatiale, à l'intonation, aux modulations d'amplitude et aux différences de F0. Les résultats de la modélisation ont montré que l'implementation de l'annulation harmonique était nécessaire pour prédire les effets liés à la F0 et leur interaction avec le démasquage spatial et l'écoute dans les trous.

Une expérience a été menée pour étudier le rôle potentiel de la suppression harmonique dans l'intelligibilité de la parole. Bien qu'il existe des preuves que l'annulation harmonique joue un rôle contre les masqueurs complexes harmoniques, son utilité contre les masqueurs naturels de parole avec des F0 non stationnaires et des parties non voisées a été étudiée. Les SRTs ont été mesurés pour sept types de masqueurs: bruit, complexes harmoniques monotonisés et intonisés, parole monotonisée, parole vocodée, parole inversée et parole naturelle. Ces stimuli ont fourni une comparaison entre les conditions avec et sans structure harmonique, modulations d'amplitude et variations de F0 au cours du temps. Cette étude a confirmé les résultats de la littérature selon lesquels, pour les complexes harmoniques, l'avantage dû à l'harmonicité est altéré par l'intonation dans le masqueur. Cette étude suggère également que la contribution de la suppression harmonique dans les situations impliquant des sources concurrentes de parole est probablement très limitée. Cette conclusion a été étayée par les prédictions du modèle avec suppression harmonique.

Ces travaux de thèse ont permis de développer un modèle d'intelligibilité de la parole pour les masqueurs complexes harmoniques. Il a également fourni des connaissances supplémentaires sur le rôle de la suppression harmonique dans les situations de cocktail party, avec des masqueurs de parole, suggérant que les effets liés à la F0 dans de telles situations pourraient être principalement dus à une réduction de l'IM. Ces travaux de thèse ont également soulevé des questions sur la potentielle interaction entre les effets liés à la F0 et le démasquage binaural. Des études supplémentaires seraient nécessaires pour mieux comprendre cela. Dans l'ensemble, ce travail suggère que les modèles d'intelligibilité de la parole validés pour les masqueurs de bruit modulés devraient être suffisants pour donner une première approximation des prédictions pour les masqueurs de parole lorsque seuls les effets d'EM sont impliqués.

References

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

Arbogast, T. L., Mason, C. R., and Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America*, 112(5):2086–2098.

Assmann, P. F. (1999). Fundamental frequency and the intelligibility of competing voices. *Proceedings of the 14th International Congress of Phonetic Sciences (San Francisco)*, pages 179–182.

Best, V., Roverud, E., Baltzell, L., Rennies, J., and Lavandier, M. (2019). The importance of a broad bandwidth for understanding "glimpsed" speech. *The Journal of the Acoustical Society of America*, 146(5):3215.

Beutelmann, R. and 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(1):331–342.

Beutelmann, R., Brand, T., and Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *The Journal of the Acoustical Society of America*, 127(4):2479–2497.

Bird, J. and Darwin, C. J. (1998). Effects of a difference in fundamental frequency in separating two sentences. In Palmer, A., Rees, A., Summersfield, Q., and Meddis, R., editors, *Psychophysical and Physiological Advances in Hearing*, pages 263–269. Wiley.

Boersma, P. and 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. and Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10(1):23–36.

Bronkhorst, A. W. and Plomp, R. (1992). Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *The Journal of the Acoustical Society of America*, 92(6):3132–3139.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3):1101–1109.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, 120(6):4007–4018.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and 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(5):2527–2538.

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(5):975–979.

Collin, B. and 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(2):1146–1159.

Conroy, C., Best, V., Jennings, T. R., and Kidd, G. (2020). The importance of processing resolution in "ideal time-frequency segregation" of masked speech and the implications for predicting speech intelligibility. *The Journal of the Acoustical Society of America*, 147(3):1648.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573.

Culling, J. F. and Darwin, C. J. (1993). Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. *The Journal of the Acoustical Society of America*, 93(6):3454–3467.

Culling, J. F., Hawley, M. L., and 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(1):552–552.

Culling, J. F. and Lavandier, M. (2021). Binaural Unmasking and Spatial Release from Masking. In Litovsky, R. Y., Goupell, M. J., Fay, R. R., and Popper, A. N., editors, *Binaural Hearing*, volume 73, pages 209–241. Springer International Publishing, Cham.

Culling, J. F. and Mansell, E. R. (2013). Speech intelligibility among modulated and spatially distributed noise sources. *The Journal of the Acoustical Society of America*, 133(4):2254–2261.

Culling, J. F. and Stone, M. A. (2017). Energetic Masking and Masking Release. In Middlebrooks, J. C., Simon, J. Z., Popper, A. N., and Fay, R. R., editors, *The Auditory System at the Cocktail Party*, volume 60 of *Springer Handbook of Auditory Research*, pages 41–73. Springer International Publishing, Cham.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 114(5):2913.

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

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(6):3271–3290.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (1997a). Concurrent vowel identification. I. Effects of relative amplitude and F0 difference. *The Journal of the Acoustical Society of America*, 101(5):2839–2847.

de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (1995). Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement. *The Journal of the Acoustical Society of America*, 97(6):3736–3748.

de Cheveigné, A., McAdams, S., and Marin, C. M. H. (1997b). Concurrent vowel identification. II. Effects of phase, harmonicity, and task. *The Journal of the Acoustical Society of America*, 101(5):2848–2856.

Deroche, M. L. D. and Culling, J. F. (2011). Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation. *The Journal of the Acoustical Society of America*, 130(5):2855–2865.

Deroche, M. L. D. and Culling, J. F. (2013). Voice segregation by difference in fundamental frequency: Effect of masker type. *The Journal of the Acoustical Society of America*, 134(5):EL465–EL470.

Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. (2014a). Roles of the target and masker fundamental frequencies in voice segregation. *The Journal of the Acoustical Society of America*, 136(3):1225–1236.

Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. (2014b). Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicity. *The Journal of the Acoustical Society of America*, 135(5):2873–2884.

Deroche, M. L. D. and Gracco, V. L. (2019). Segregation of voices with single or double fundamental frequencies. *The Journal of the Acoustical Society of America*, 145(2):847–857.

Durlach, N. I. (1972). Binaural signal detection - Equalization and cancellation theory. Technical report.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (2003). Note on informational masking (L). *The Journal of the Acoustical Society of America*, 113(6):2984–2987.

Ewert, S. D., Schubotz, W., Brand, T., and Kollmeier, B. (2017). Binaural masking release in symmetric listening conditions with spectro-temporally modulated maskers. *The Journal of the Acoustical Society of America*, 142(1):12–28.

Festen, J. M. and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88(4):1725–1736.

Fogerty, D., Xu, J., and Gibbs, B. E. (2016). Modulation masking and glimpsing of natural and vocoded speech during single-talker modulated noise: Effect of the modulation spectrum. *The Journal of the Acoustical Society of America*, 140(3):1800–1816.

Freyman, R. L., Balakrishnan, U., and 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(5):2246–2256.

Gardner, B. and Martin, K. (1994). HRTF Measurements of a KEMAR Dummy-Head Microphone. *MIT Media Lab Percept. Comput.*, pages 1–7.

Glasberg, B. R. and Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1):103–138.

Green, T. and 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(4):2876–2883.

Guest, D. R. and Oxenham, A. J. (2019). The role of pitch and harmonic cancellation when listening to speech in harmonic background sounds. *The Journal of the Acoustical Society of America*, 145(5):3011–3023.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115(2):833–843.

Jackson, H. M. and Moore, B. C. J. (2013). Contribution of temporal fine structure information and fundamental frequency separation to intelligibility in a competing-speaker paradigm. *The Journal of the Acoustical Society of America*, 133(4):2421–2430.

Jelfs, S., Culling, J. F., and Lavandier, M. (2011). Revision and validation of a binaural model for speech intelligibility in noise. *Hearing Research*, 275(1-2):96–104.

Jørgensen, S. and Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America*, 130(3):1475–1487.

Jørgensen, S., Ewert, S. D., and Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1):436–446.

Keysers, C., Gazzola, V., and Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, 23(7):788–799.

Kidd, G., Best, V., and 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(6):3793–3802.

Kidd, G. and Colburn, H. S. (2017). Informational Masking in Speech Recognition. In Middlebrooks, J. C., Simon, J. Z., Popper, A. N., and Fay, R. R., editors, *The Auditory System at the Cocktail Party*, volume 60 of *Springer Handbook of Auditory Research*, pages 75–109. Springer International Publishing, Cham.

Kidd, G., Mason, C. R., and Gallun, F. J. (2005). Combining energetic and informational masking for speech identification. *The Journal of the Acoustical Society of America*, 118(2):982–992.

Kidd, G., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., and Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America*, 140(1):132–144.

Kwon, B. J. and Turner, C. W. (2001). Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference?a). *J. Acoust. Soc. Am.*, 110(2):11.

Lavandier, M. and Best, V. (2020). Modeling Binaural Speech Understanding in Complex Situations. In Blauert, J. and Braasch, J., editors, *The Technology of Binaural Understanding*, Modern Acoustics and Signal Processing, pages 547–578. Springer International Publishing, Cham.

Lavandier, M. and Culling, J. F. (2010). Prediction of binaural speech intelligibility against noise in rooms. *The Journal of the Acoustical Society of America*, 127(1):387–399.

Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., and 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(1):218–231.

Leclère, T., Lavandier, M., and 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.

Miller, G. A. and Licklider, J. C. R. (1950). The Intelligibility of Interrupted Speech. *The Journal of the Acoustical Society of America*, 22(2):167–173.

Patterson, R., Nimmo-Smith, I., Holdsworth, J., and 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.*, page 34.

Peters, R. W., Moore, B. C. J., and Baer, T. (1998). Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *The Journal of the Acoustical Society of America*, 103(1):577–587.

Plomp, R. (1976). Binaural and Monaural Speech Intelligibility of Connected Discourse in Reverberation as a Function of Azimuth of a Single Competing Sound Source (Speech or Noise). *Acta Acustica united with Acustica*, 34(4):200–211.

Popham, S., Boebinger, D., Ellis, D. P. W., Kawahara, H., and McDermott, J. H. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature Communications*, 9(1):2122.

Prud'homme, L., Lavandier, M., and 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(5):3246–3254.

Rennies, J., Best, V., Roverud, E., and Kidd, G. (2019). Energetic and Informational Components of Speech-on-Speech Masking in Binaural Speech Intelligibility and Perceived Listening Effort. *Trends in Hearing*, 23:2331216519854597.

Rhebergen, K. S. and Versfeld, N. J. (2005). A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 117(4):2181–2192.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *The Journal of the Acoustical Society of America*, 120(6):3988–3997.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246.

Schoenmaker, E., Brand, T., and van de Par, S. (2016). The multiple contributions of interaural differences to improved speech intelligibility in multitalker scenarios. *The Journal of the Acoustical Society of America*, 139(5):2589–2603.

Schoof, T. and Rosen, S. (2015). High sentence predictability increases the fluctuating masker benefit. *The Journal of the Acoustical Society of America*, 138(3):EL181–EL186.

Shackleton, T. M. and Carlyon, R. P. (1994). The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *The Journal of the Acoustical Society of America*, 95(6):3529–3540.

Simpson, S. A. and Cooke, M. (2005). Consonant identification in N-talker babble is a non-monotonic function of N. *The Journal of the Acoustical Society of America*, 118(5):2775–2778.

Steinmetzger, K. and 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(6):3586–3599.

Steinmetzger, K., Zaar, J., Relaño-Iborra, H., Rosen, S., and 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(4):2562–2576.

Stone, M. A. and Canavan, S. (2016). The near non-existence of "pure" energetic masking release for speech: Extension to spectro-temporal modulation and glimpsing. *The Journal of the Acoustical Society of America*, 140(2):832–842.

Stone, M. A., Füllgrabe, C., Mackinnon, R. C., and 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(5):2874–2881.

Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America*, 132(1):317–326.

Stone, M. A. and Moore, B. C. J. (2014). On the near non-existence of "pure" energetic masking release for speech. *The Journal of the Acoustical Society of America*, 135(4):1967–1977.

Summerfield, Q. and Assmann, P. F. (1991). Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony. *The Journal of the Acoustical Society of America*, 89(3):1364–1377.

Summerfield, Q. and Culling, J. F. (1992). Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency. *The Journal of the Acoustical Society of America*, 92(4):2317–2317.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

Vicente, T. and Lavandier, M. (2020). Further validation of a binaural model predicting speech intelligibility against envelope-modulated noises. *Hearing Research*, 390:107937.

Vicente, T., Lavandier, M., and 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(5):3305–3317.

Watson, C. S. (2005). Some Comments on Informational Masking. *Acta Acustica united with Acustica*, 91(3):502–512.

Westermann, A. and Buchholz, J. M. (2015). The influence of informational masking in reverberant, multi-talker environments. *The Journal of the Acoustical Society of America*, 138(2):584–593.

Wichmann, F. A. and Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8):1293–1313.