Effect of contralateral noise on energetic and informational masking on speech-in-speech intelligibility

Marjorie Dole\textsuperscript{1}, Michel Hoen\textsuperscript{2}, Fanny Meunier\textsuperscript{1}.

\textsuperscript{1} Laboratoire Dynamique du Langage, CNRS UMR 5596, Lyon, France
\textsuperscript{2} Stem Cell and Brain Research Institute, INSERM U846, Lyon, France
marjorie.dole@univ-lyon2.fr

Abstract

This experiment tested the advantage of binaural presentation of an interfering noise in a task involving identification of monaurally-presented words. These words were embedded in three types of noise: a stationary noise, a speech-modulated noise and a speech-babble noise, in order to assess energetic and informational masking contributions to binaural unmasking. Our results showed important informational masking in the monaural condition, principally due to lexical and phonetic competition. We also found a binaural unmasking effect, which was more important when speech was used as interferer, suggesting that this suppressive effect was more efficient in the case of high-level informational (lexical and phonetic) competition.

Index Terms: cocktail party, speech-in-speech, speech-in-noise, binaural unmasking, informational masking

1. Introduction

In everyday social life, we are often confronted to situations in which we must segregate speech sounds from concurrent noise. Inside the general context of the cocktail party effect, one particular situation can be of first relevance for who is interested in lexical access and speech comprehension, namely speech-in-speech comprehension.

Two complementary types of masking phenomena occur in cocktail party situations: energetic masking (EM), which is produced when the target signal and interfering noise partly overlap in time and frequency at peripheral level, and informational masking (IM) which occurs when information inside the target signal and the interfering noise is of comparable nature and become difficult to disentangle [1]. IM is thought to operate at higher level than the periphery. Typically, IM occurs when the competing flows are speech sounds. In such situations, IM gets particularly important and made up of a complex of different levels of linguistic information, creating competitions between lexical, phonetic, and semantic cues. During speech-in-speech comprehension, both energetic and informational masking occur, although it has been shown that energetic masking represents only a small part of the overall amount of masking [2].

In the speech-in-speech situation, segmentation between competing sounds can be facilitated by different cues, one of them being for example fundamental frequency (F0): when two competing voices are presented, a difference in their F0 improves their intelligibility. Another well characterized cue is spatial location: when the target signal and competing noise originate from the same location, noise greatly affects the intelligibility of the target; but when the target signal and the competing noise are spatially separated, the intelligibility of the target increases. This is reflected in the binaural release of masking effect. It refers to the observed improvement of intelligibility of a target when noise is added to the contralateral ear while target and noise are presented to the ipsilateral ear, in comparison to the simple monaural situation in which noise and target are presented to the same ipsilateral ear [3],[4] and [5]. This effect relies on the fact that binaural cues help differentiating the two competing flows.

The relationship between informational/energetic masking and binaural release of masking is still only partly understood. Does binaural unmasking affect energetic masking and informational masking in the same way? How is this effect modulated by the nature of the concurrent noise? Recent studies have shown that binaural release of masking was more efficient in the case of informational masking [6]; for example, Hawley \textit{et al.} [7] tested four types of noise: speech, reversed speech, speech-shaped noise, and speech-modulated noise, with interferers either spatially coincident with or separated from the target. They measured binaural advantage and found for two or three interferers a 2–4dB binaural advantage for noise and speech-modulated noise, and 6–7dB for speech and time-reversed speech, suggesting that binaural release of masking is more important when multiple interferers are voiced, or when IM is more important.

Since binaural release of masking appears to be more important in informational masking, it seems interesting to investigate further the complex informational interactions that can be produced by concurrent speech in speech-in-speech comprehension. Obviously, speech is a complex aggregate of different levels of information that are all potential sources of IM. What levels of linguistic information participate to IM and which ones are more sensitive to binaural release of masking?

We started addressing this issue in a previous study on the characterisation of psycholinguistic levels inside the IM [8], using a diotic configuration of presentation. We used 4 to 8-talkers babble and reversed-babble sounds, in order to investigate lexical interferences that could exist with these concurrent sounds. We have shown that the informational masking occurring in speech in speech comprehension situations consisted of different levels of competitions, including an acoustic-phonetic competition (which was similar in natural and reversed speech) and a lexical competition reflected in the increase of masking when noise was natural speech in contrast with reversed speech. We also found that this effect was modulated by the number of talkers in the cocktail noise, lexical competition being present only in a babble made up of 4-voices, leaving higher order cues available, but disappeared with increasing number of talkers, as the signal progressively saturated and higher-level cues got less available.

The purpose of the present study was to further investigate these informational masking effects in the context of binaural unmasking. In order to separate different levels of information participating to the informational unmasking caused by speech and determine their sensitivity to binaural unmasking, we used...
three noises: a 4-talkers cocktail (Cocktail) constituting a high-
level informational masker, a speech-modulated noise (SN) which
contained the same long-term average spectrum as the
babble but no temporal fine structure or any phonemic
information, and a stationary broadband noise (BBN)
producing only energetic masking. These noises were tested in
three configurations of presentation: a monaural one (target
and interfering noise presented in the same ear), a pure
dichotic configuration (target in one ear and noise in the other
ear); and a binaural configuration (target in one ear and noise
in two ears).

2. Methods

2.1. Participants and procedure

76 volunteers participated in this experiment. All were right
handed native French speakers, aged 18-35 years, and had no
known hearing or language disorders.

They were asked to listen to auditory stimuli, delivered via
headphones (Beyerdynamic DT48, 200 Ω). Stimuli were mono-
or bisyllabic words, presented in different types of
noises. Subjects were asked to handwrite the word they heard.

2.2. Stimuli

2.2.1. Target words

126 monosyllabic and 126 bisyllabic words were selected in a
middle range of frequency of occurrence (ranging from 0.23 to
364.26, mean: 20.94, S.D.: 41.86 for monosyllabic words, and
from 0.23 to 338.19, mean: 16.81, S.D.: 43.74 for bisyllabic
words) according to the French database Lexique2. [9] They
were pronounced by a 24-year-old French native female
speaker and recorded in a sound-proof room.

2.2.2. Noises

Three types of noises were used: a Cocktail noise (Cocktail), a
fluctuating speech-shaped noise (SN) and a stationary
broadband noise (BBN).

Cocktail noise was made up of 4 voices (2 male; 2 female);
every single voice was recorded in a sound-proof room,
reading extracts of French press. Individual recordings were
modified according to the following protocol: 1) removal of
silence and pauses of more than 1 s; 2) suppression of
sentences containing pronunciation errors, exaggerated
prosody or proper nouns; 3) noise reduction optimized for
speech signals, 4) intensity calibration in dB-A and
normalization of each source at 80dB-A; 5) final mixing of
individual sources into cocktail party sound tracks.

The fluctuating speech-shaped noise (SN) was made to have spectro-temporal characteristics comparable to that of our
cocktail noise; to do so, we took our 4-voices cocktail noise and extracted envelope information below 60Hz. Using Fast
Fourier Transformation (FFT), the power spectrum and phase
distribution of the original signal were computed and original
phase information discarded by randomizing phase
distribution. An inverse FFT was used to generate a new signal
with equivalent power spectrum but randomized phases
convolved with the temporal envelope of the original cocktail
noise. Finally, the root mean square (rms) powers of the
original and new signals were matched.

Finally, a stationary broadband noise with spectral
composition identical to our cocktail noise was generated
(BBN).

2.3. Stimuli and word lists

Stimuli consisted in a block of 126 monosyllabic words and a
block of 126 bisyllabic words. Stimuli were composed by mixing each target word with randomly selected 4s chunks of
noise, target words were inserted 2.5s after the onset of the
noise.

3 different presentation configurations were tested: i) a
dichotic configuration, with the target word in ipsilateral ear
and noise in contralateral ear; ii) a monaural configuration,
with the target word and the noise in the same ear and iii) a
hybrid dichotic/monaural configuration, with target word in
ipsilateral ear, and the same noise in the two ears.

For every configuration of presentation, the 3 noises were
tested leading to 9 conditions, and 14 words per condition.

Table 1. Conditions used in this experiment

<table>
<thead>
<tr>
<th>Condition</th>
<th>Target ear</th>
<th>Contralateral ear</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_N</td>
<td>Speech</td>
<td>Noise</td>
<td>BBN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cocktail</td>
</tr>
<tr>
<td>SN_Si</td>
<td>Speech/noise</td>
<td>Silence</td>
<td>BBN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cocktail</td>
</tr>
<tr>
<td>SN_N</td>
<td>Speech/noise</td>
<td>Noise</td>
<td>BBN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cocktail</td>
</tr>
</tbody>
</table>

Stimuli were presented at a SNR of 0dB in the target ear and
an intensity of 20dB lower in the contralateral ear in
comparison with target ear (ILD=-20dB).

18 lists of stimuli were created (9 lists composed with
monosyllabic stimuli and 9 lists with bisyllabic stimuli), so
that across the lists each word was presented in every
condition and every noises. Within lists, frequency of words
was counterbalanced.

Each subject heard two lists of stimuli, called ‘blocks’; half of
the subjects listened to a block of monosyllabic words in the
right ear and a block of bisyllabic words in the left ear, and the
other half of subjects listened to a block composed of
monosyllabic words in their left ear and a block of bisyllabic
words in their right ear. Order of presentation of blocks was
randomized.

3. Results

Subjects’ responses were analysed by calculating the
proportion of words that corresponded to the target word.

These individual word identification rates were then used as
dependant variables in the following analysis. A 4-way
repeated-measures ANOVA was performed, with target ear (Ear)
and word length (Length) as inter-subject factors, and
noise (Noise) and presentation configuration (Configuration)
as intra-subject factors.

This analysis revealed a significant effect of Noise (Figure
1, F (2, 296) = 40.602, p < .001); planned comparisons
revealed that all noises differed from each other, with Cocktail
noise producing more errors than SN (F (1, 75) = 26.819, p <
.001), which produces more errors than BBN (F (1, 75) =
8.845, p < .005).

We found also a significant effect of Configuration (Figure
2, F (2, 232) = 1654.323, p < .001). Planned comparisons
revealed that all conditions differed: in S_N configuration,
subjects obtained better scores (98% of correct responses) than
in SN_Si (F (1, 75) = 1350.752, p < .001) which produced
64.4% of correct responses; adding a contralateral noise in SN_N improved performances (74.4%) in comparison with SN_Si configuration (F (1, 75) = 88.684, p< .001). Performances in S_N and SN_N configurations were also significantly different (F (1, 75) = 1409.087; p< .01).

Figure 1: Correct identification rate as a function of type of noise. Error bars represent ± 1 S.D

There was a significant Noise*Configuration interaction (F (4,592) = 28.057, p< .001), as shown on figure 2, revealing that in SN_Si condition, Cocktail produced significantly more interference than SN (F (1, 75) = 57.13, p< .001), which produced more interference than BBN (F (1, 75) = 5.46, p< .05). In SN_N condition, BBN produced significantly less errors than SN (F (1, 75) = 6.20, p< .05) and Cocktail (F (1, 75) = 4.41, p< .05).

We also observed a significant effect of Length: bisyllabic words were globally better recognized (83.2%) than the monosyllabic ones (74.8%) (F (1, 148) = 136.317, p< .001).

Moreover, the ANOVA revealed an interaction of Length*Configuration (F (2, 296) = 48.845, p< .001). However this seemed to be explained by a ceiling effect in the S_N condition: planned comparisons revealed that in S_N condition performances were not significantly different between monosyllabic and bisyllabic words (F (1,150) = 0.625; p = 0.43). This difference between mono and bisyllabic words became significant for SN_Si (F (1, 150) = 56.89; p< .001) and SN_N (F (1,150) = 128.13; p< .001) configurations.

Finally, there was a significant effect of Ear (F (1, 148) = 4.407; p < 0.05, with better intelligibility when target items were presented in the left ear (79.8%) than when they were presented in the right ear (78.2%).

Binaural unmasking effect was evaluated by subtracting performances obtained in SN_Si configuration from performances obtained in SN_N configuration. (Figure 3). A one-way ANOVA was then effectuated, with Noise as within factor. We obtained a significant effect (F (2, 150) = 26.47, p< .001). Planned comparisons revealed that BBN and SN didn’t differ (F (1, 75) = 0.268, p = 0.606), whereas cocktail differed from both BBN (F (1, 75) = 32.02, p< .001) and SN (F (1, 75) = 42.35, p< .001).

Figure 3: Percent of masking release in the SN_N configuration. This was obtained subtracting correct identification rate in SN_Si configuration from ones obtained in SN_N configuration

4. Discussion

In this study we were interested in speech comprehension in the presence of concurrent speech and we were focused on the unmasking effect which occurs when contralateral noise is added to a monaural situation where noise and target are presented to the same ear.

Our results showed that in a pure dichotical configuration performances ceiling near 100% of correct identification, and this for all types of noise used. This was in agreement with
results observed in the literature [10], [11] and it can be explained by the fact that masking is greatly reduced in this configuration, due to maximal separation of the two competing sources.

In the monaural condition, competing noises were presented to the same ear as target words; this produced an important masking effect, with performances decreasing from 100% to 60% on average. This can be explained by the fact that noise and target were not spatially separated, and segregation between target and interferer became difficult. In this condition we had a significant effect of the type of noise: As expected, the BBN produced less interference (70% of correct responses) which can be explained by the fact that this noise contained neither linguistic information nor temporal fluctuations and therefore produced only stationary energetic masking. Cocktail was the most interfering noise, and decreased performances dramatically, down to 57%. This was in agreement with results previously observed [12]; in this case, target and noise shared the same frequency, the same temporal fluctuations in their spectrum and the same phonological and lexical components. This resulted in a basic energetic masking, plus important high-level informational masking.

Speech shaped noise constitutes an intermediate situation, with performances reaching an average of 68% in this configuration; this noise contained the same temporal envelope as cocktail masker, but the fine temporal structure was suppressed. This could be considered as a first-level informational masking but with minimal magnitude here (2%). The fact that Cocktail degrades performances in a more important manner than SN suggests that informational masking greatly depends on higher-level informational cues like temporal fine structure, phonological and lexical information.

When noise was presented diotically and target monaurally, we observed an important unmasking effect causing an increase of performance in this configuration in comparison with the monaural configuration. This can be explained by the perceived location effect: when binaural cues are available, they help the listener to segregate between target and noise. It should be noted that in this experiment, only ILD was applied to contralateral noise; the fact that we observed an important masking release suggest that it constitutes a sufficient cue to allow a significant spatial masking release. This can be explained by the fact that binaural cues are available to help the listener to segregate between target and noise. In this configuration, we had a significant effect of Noise, with BBN producing less interference than SN and Cocktail, but no difference between SN and Cocktail, suggesting that the increased informational masking observed for babbble noise in the monaural configuration was cancelled by binaural release of masking.

As a result, the magnitude of binaural release of masking was larger for Cocktail than for the two other noises, suggesting that informational unmasking is more important than energetic masking release (Figure 3).This was in agreement with results obtained by Arbogast & al. [6]. This study also tested the effect of word length; we obtained better performances for bisyllabic words than for monosyllabic words, in SN_Si and SN_N conditions. This could be explained by two different mechanisms: redundancy of acoustical cues: the first syllable contained cues that facilitate recognition of the second syllable. The second mechanism refers to the fact that when listeners got the first syllable they could guess the second syllable in order to obtain the word.

5. Conclusions

The aim of this study was to investigate energetic and informational masking release that occurs when noise is added in a binaural manner to a monaural target. We were principally interested in evaluating the fine cues that are important to cause an informational masking. Our results showed that high level linguistic cues like lexical and phonetic information are predominant inside informational masking and are also very susceptible to binaural release of masking whereas lower-level information sources as spectral of envelope information cause minor informational masking effects and are also less susceptible to be cancelled by binaural unmasking.

6. Acknowledgements

The first author’s thesis work is funded by a PhD grant from the “Cluster régional de recherche: Handicap Vieillissement et Neurosciences (HVN)” from the Rhône-Alpes region. We are grateful to the European Research Council for funding this project via an ERC starting grant attributed to Fanny Meunier (SpiN project).

7. References