Who hides behind the masks at the cocktail party?

Behavioral assessment of different types of interferences occurring during speech-in-noise comprehension

Michel Hoen1,2, Claire-Léonie Grataloup1, Nicolas Grimault2, Fabien Perrin2,

François Pellegrino1, Fanny Meunier1, Lionel Collet2

1Laboratoire dynamique du Langage
UMR5096 CNRS, Université Lumière, Lyon, France

2Laboratoire Neurosciences et Systèmes Sensoriels
UMR5020 CNRS, Université Claude Bernard, Lyon, France

michel.hoen@phonak.ch

ABSTRACT

Up to now speech in noise comprehension and more particularly speech in concurrent speech sounds was rarely studied in the domain of psycholinguistics. In this paper we report a study interested in the differential effects of different types of speech derived noises as multi-talker cocktail party sounds and their time-reversed pendent on the comprehension of isolated words. Results suggest that different levels of linguistic information from concurrent speech signals can compete with linguistic information in the target signal, mainly depending on the spectral saturation caused by the increasing number of voices in concurrent signals. These results suggest linguistically specific participations in informational masking effects occurring in the context of speech in speech comprehension.

1. INTRODUCTION

Although we are of course quite good at understanding speech delivered via headphones in a silent anechoic experimental room, we are more usually confronted to the situation when speech occurs in the acoustic chaos of a noisy crossroad or in the babbling crowd of a reunion room and still quite well able to understand the messages it transmits. Speech often remains intelligible under extremely variable acoustical conditions, despite important types and amounts of possible interfering background noises, reverberations or concurrent discussions. This phenomenon, referred to as the 'cocktail party effect' consists of a specialized cognitive ability to selectively focus our attention on one particular auditory flow amongst one or more concurrent flows.

Since its first description by Cherry in 1953 [1], the cocktail party effect gave rise to numerous experiments that helped characterizing the cognitive basics of this unique ability. It has been shown for example that spatial segregation of sounds constitutes one of the main cues allowing the cocktail party effect [2]. However, when the different target and concurrent sound flows originate from one identical spatial location, or in the case of diotic sound presentation of competing speech signals, the cognitive system must rely on a mosaic of other cues in order to isolate the signal of interest. Former studies have provided evidence that different acoustical cues could help in discriminating the target voice from concurrent voices, as individual characteristics (fundamental frequency, accent, discursive style etc.), prosodic modulations or global intensity differences existing between the different sources [3-8].

In diotic cocktail party situations, effects of interfering signals were described as potentially masking target signals at at least two differentiable levels, corresponding to two different masking effects, namely 'energetic masking' and 'informational masking' [7-8]. Energetic masking is due to the spectrotemporal composition of concurrent sounds and occurs whenever a speech signal is emitted in the presence of a broadband noise that superimposes at least partially in time and frequency to the target signal. Informational masking is attributed to the type of information present in the interfering noise [4-5]. In this case physical signals of the target and concurrent sounds do not have to overlap necessarily, but the information they contain compete during later analysis stages. Informational masking is particularly salient in the context of speech in speech comprehension. In this context, Brungart et al., studied the effects of voice number, and signal-to-noise ratio (SNR) in situations with 1, 2 and 3 concurrent voices [7-8]. Their results showed a linear decrease of performances with SNRs in situations where more than one concurrent voice was present. However, although the study of the cocktail party effect gave rise to an impressive number of studies, the detailed linguistic interferences occurring in speech in speech situations were rarely questioned if ever. In particular, informational masking in speech in speech situations could be attributed to various levels of linguistic information as phonologic, lexical or prosodic cues and one could imagine the existence inside the informational masking effect of various subtypes of linguistic masks. The determining of these different types of masking effects of course is related to the potential availability, inside the concurrent signals of such information as phonological or lexical information and this availability potentially depends on the number of voices present in the concurrent noise, as well as on the SNR at which target and masking signals are presented.

The current experiment was designed to study the possible existence of different levels of interfering linguistic information in multi-talkers speech in speech comprehension situations. We therefore ran a single-word comprehension experiment performed in concurrent noise conditions in which we manipulated both the number of voices implicated in the concurrent noise and the type of concurrent noise. We used signals composed of 4, 6 and 8 voices and the same speech sounds but time-reversed (reversed cocktails). Time reversion had sometimes been argued to constitute the most drastic degradation one could apply to speech [9]. However, not only does reversed speech 'sound' like speech, it also keeps some of the linguistic information present in natural speech, as vowels for example. Moreover, when different reversed-speech flows are mixed together, this speech noise sounds as speech and one can clearly perceive phonemes in it, but of course no words. To differentiate pure energetic masking from informational masking, we also included a broadband noise, with spectro-temporal characteristics similar to our speech noises.

2. MATERIALS AND METHODS

2.1. *Interfering Sounds*

Multi-talker cocktail party sounds

Three cocktail party sounds were created by mixing 4, 6 or 8 single speaker sources. Only one voice was active at each point in time in each individual track. Each speaker was first recorded separately in a sound-proof room using a Røde NT1 microphone, an Ultragain MIC 2000 pre-amplifier and a RolandED UA-30 soundcard, at a sampling frequency of 44 kHz, 16 bits. Each original source was composed of intelligible French sentences, made free of proper nouns and modified according to the above mentioned protocol: i) suppression of silences and pauses superior to 1 second, ii) suppression of sentences containing pronunciation errors, prosodic overplay or proper nouns, iii) noise reduction optimized for speech signals (CoolEdit Pro© 1.1 – Dynamics Range Processing – preset Vocal limiter), iv) intensity calibration in dB-A and normalization of each source at 80 dB-A (Larson Davis system LD824, artificial ear: AEC101 acoustic coupler, Sennheiser HD250 linear II headphones) and v) final mixing of individual sources into cocktail party sound tracks, saved as .wav files (44kHz, 16 bits, Stereo).

Reversed Cocktail party sounds

Reversed cocktail party sounds were obtained by reversing the previously generated cocktail party sounds along their temporal dimension using the inbuilt tool of the software Adobe® Audition™ in its 1.0 version.

Associated broadband noise

In order to obtain a broadband noise with comparable spectrotemporal characteristics to our speech and reversed speech sounds, we used the 8 speaker cocktail party noise as reference, as this noise was supposed to potentially exhibit the strongest energetic masking effect because it has both the most complex spectral composition of all speech sounds and statistically the most constant temporal envelope due to the higher number of slow time-varying acoustical flows it contains. We then extracted its envelope below 60Hz, in order to obtain the corresponding shape of slow temporal variations it contained. Then, using Fast Fourier Transformation (FFT), we computed the power spectrum and phases distribution from the original signal and discarded original phase information by randomizing phase distribution. By then applying an inverse FFT to the resulting signal, we obtained a new signal with equivalent power spectrum but randomized phases that was convolved with the temporal envelope of the original cocktail party sound. Finally, the rms power of the new signal was determined and adjusted to the rms power of the original cocktail party sound. That way, we obtained a broadband noise, with both spectral composition and slow temporal variations comparable to that of our 8 speaker cocktail party sound, but that contained no linguistic information anymore.

2.2. *Target words*

We recorded a list of 320 French mono-syllabic, tri-phonemic words. They were selected in a middle frequency range (from: 0.19 pm, to: 146.71 pm; average = 20.96; SD = 21.37) according to the French database Lexique2 [10], to avoid extremely high or low frequency items that subjects typically overuse or ignore. Single words were pronounced by a single male native French speaker in a sound-proof cabin using a Sony ECM-MS907 microphone and saved as .wav files (44 kHz, stereo, 16 bits). Words were manually segmented from surrounding noise using the software Adobe® Audition™ in its 1.0 version and acoustical quality of recordings ensured by keeping word's signal to surrounding noise ratio over at least 30dB.

2.3 *Conditions randomization and Subjects*

In total, we tested 9 concurrent noise types:

- 4, 6 and 8 voices (mixed genders) cocktail party
- 4, 6 and 8 voices (mixed genders) reversed cocktail party
- physical noise associated to the 8 voices (mixed genders) cocktail party sound

These noises were all tested at 4 signal-to-noise ratios of -3, 0, +3 and +6 dBs, yielding a total of 36 conditions. Hence, we generated 36 lists of 8 words, matched for frequency and number of phonological neighbours, words in final experimental lists had an average frequency of 3.92 pm (SD=0.006) and an average number of phonological neighbours of 19.83 (SD=0.06). 36 subjects participated in the experiment; they were all students, aged 18 to 32 years and native French speakers with no known hearing or language disorders. Participants were paid for their participation.

Stimuli

Stimuli consisted of 288 .wav files (44 kHz, 16bits, stereo) each lasting for 4 sec. Background noise was present during the whole stimulus and the target word was inserted always at 2.5 sec from the start. Target words had durations from 242 ms to 914 ms (mean = 550 ms, SD = 135 ms). Stimuli were composed by mixing one chunk of noise, randomly selected amongst 40 chunks of 4 sec extracted from the original noise files with on word, at a particular signal-to-noise ratio. Individual intensity levels for noise and target word were fixed according to the global rms power of both original sounds to mix. As the mixing procedure results in some global intensity modulation and in order to avoid that the global intensity of stimuli would be predictive of the SNR in the stimuli, a final randomized intensity roving over a $^{+}$/. 3 dB range in 1 dB steps was applied to every created stimuli.

Procedure

Subjects sat in a quite room facing a computer monitor. Stimuli were delivered dioticaly via headphones (Beyerdynamic DT 48, 200Ω) at a comfortable level. All noises, word lists and conditions were randomized across subjects. Subjects were informed that they would here words pronounced by one single male speaker against different types of background noises. The task consisted in a single word transcription task, subjects being asked to type the sounds they heard on a computer keyboard. Subjects were asked to type words if they could or to transcript the sound they heard when possible. Before the testing phase, subjects were proposed 12 training items so they could accommodate to the stimulus presentation mode and target speaker's voice. Timing of the experiment was determined by the subjects themselves and total duration of the experiment was of 45 minutes in average, ranging from 30 to 60 minutes depending on participant's typing habits.

Behavioral performance assessment

Subject's transcriptions were analysed in terms of proportion of lexical retrieval, by calculating the proportion of words in the subject's transcriptions that actually corresponded to target words.

3. Results

In order to assess the effect of number of voices present in concurrent speech signals, we performed a 3-way Anova on individual lexical retrieval rates considered as dependent variable and including the factors noise Type (2: cocktail, reversed cocktail); Number of voices (3: 4, 6 and 8) and SNRs (4: -3, 0, +3 and +6 dB). This analysis revealed a significant main effect only for factor SNRs (F(3,105)= 185,31; p<.05) lexical retrieval rates decreasing linearly with SNR in our SNRs range. Main effects of factors Type (F(1,35)= 1,18; n.s.) and Number (F(2,70)= 2,19; n.s) remained non significant, suggesting an absence of global effect of both factors. However, the effects of factors Type and Number interacted significantly (F(2,70)= 4,53; p<.05). No other interactions were significant, in particular the factor SNRs did not interact significantly with any other factor, neither Type (F<1), nor Number (F<1), suggesting that lexical retrieval rates decreased linearly with SNR in our SNRs range in every experimental condition considered. Therefore, we decided to discard the SNR factor from further analyses by averaging all behavioural data over the different SNRs.

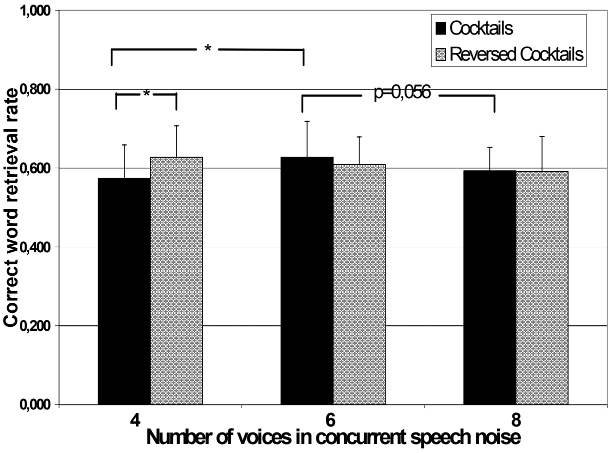

Figure 1: Behavioral results observed in the single word comprehension task as a fun ction of speech noise type.

Planned comparisons (LSD test α=.05) were performed on the second level interaction between factors Type and Number (see Fig.1). Between cocktail noises, this analysis revealed a significant difference between the cocktail with 6 voices and the cocktail with 4 voices (p=0.004) and a marginally significant difference between the cocktail with 6 voices and the cocktail with 8 voices (p=0.056), the masking effect due to the cocktail with 6 voices being lower than the masking effect due to the cocktails with 4 and 8 voices. Planned comparisons also revealed a significant difference between the cocktail party with 4 voices and the reversed cocktail party with 4 voices (p=0.004), the cocktail party with 4 voices exhibiting a stronger masking effect than the same but time-reversed noise. This was not the case for the cocktail party with 6 and 8 voices. Although the effect of masking due to reversed cocktail noises seems to increase linearly with the number of voices they are composed of, this trend leads to very small and variable differences and therefore no significant differences were observed between the effects of masking due to the different reversed-cocktail noises.

In the former analysis, we observed a significant difference between the effect of masking due to the cocktail party noise with 4 voices and the effect of masking due to the same but time-reversed noise. This observation would suggest the existence of two different informational masking effects in the 4 voices speech noise condition, only if both of these masking effects remain different from the pure energetic masking effect due to a spectro-temporally comparable physical noise. In order to test this hypothesis we finally performed a 1-way Anova, first considering lexical retrieval rates as the dependent variable, discarding SNR information and including one single factor Type (3: Associated broadband noise; 4 voices cocktail party noise; 4 voices reversed cocktail party noise). This analysis revealed a main effect of factor Type (F(2,70)=23,45; p<.05) and planned comparisons revealed that all individual differences were significant with all p values below .002, confirming that all three noise types had different masking effects on word comprehension rates, broadband noise exhibiting a smaller masking effect than reversed cocktail, the latter exhibiting a smaller effect than the speech cocktail noise (Fig.2).

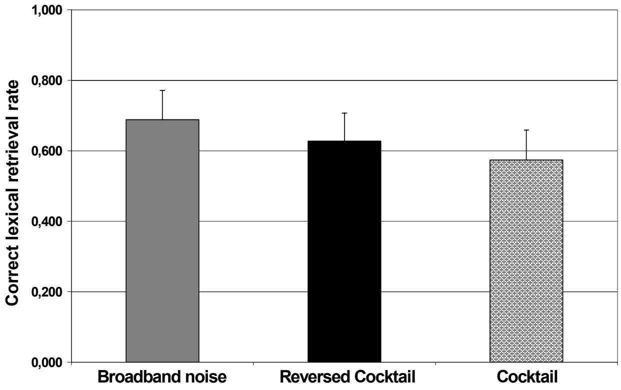

Figure 2 : Behavioral results observed in the single word comprehension task for three noise types: the speech associated broadband noise, the reversed 4 voice cocktail noise and the 4 voices cocktail noise.

4. Discussion and Conclusion

Results from this multi-talker speech in noise comprehension experiment testing the effects of the number of concurrent voices in a cocktail party situation demonstrate differential specific effects of linguistic information in cocktail party situations. We have shown that a physical broadband noise, 4 voices reversed cocktail party and 4 voices cocktail party sound had increasing masking effects identifying three levels of masking in this condition. These effects would be attributed energetic, phonologic and lexical masking levels. Indeed, the difference between reversed speech cocktail and speech cocktail at 4 voices level is that the only the latter one contains intelligible words, the first one containing only phonemes. This added lexical masking effect disappears over 4 voices, when words are no more intelligible in the cocktail party sounds because of information saturation. At 6 and 8 voices number, only the phonologic masking effect is operant, and it is increasing with saturation from 6 to 8 voices. Lexical access can be saturated by increasing the amount of phonemic information competing for lexical representation.

Moreover our experiment demonstrate that the cocktail paradigm can give the opportunity to psycholinguists to study lexical competition in a better on line situation than priming paradigm actually in use in this type of investigation.

Acknowledgements

This study was supported by an ACI (n° 67068) of the French *Ministère de l'Enseignement Supérieur et de la Recherche* attributed to Fanny Meunier.

Bibliographie

[1] E. Cherry. Some experiments on the recognition of speech, with one and two ears, *J. Acoust. Soc. Am.* 25,975–979, 1953.

[2] A. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions, Acustica 86, 117–128, 2000; *Trans. Speech and Audio Proc.*, 7(6):697-708, 1999.

[3] J. Egan, E. Carterette, and E. Thwing. Factors affecting multi channel listening, *J. Acoust. Soc. Am.* 26, 774–782, 1954.

[4] D. Dirks, and D. Bower. Masking effects of speech competing messages, *J. Speech Hear. Res.* 12, 229–245, 1969.

[5] J. Festen, and R. Plomp. Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing, *J. Acoust. Soc. Am.* 88, 1725–1736, 1990.

[6] C. Darwin, and R. Hukin. Effectiveness of spatial cues, prosody and talker characteristics in selective attention, *J. Acoust. Soc. Am.* 107, 970–977, 2000.

[7] D. Brungart. Informational and energetic masking effects in the perception of two simultaneous talkers, *J. Acoust. Soc. Am.* 109, 1101– 1109, 2001a.

[8] D. Brungart. Informational and energetic masking effects in the perception of multiple simultaneous talkers, *J. Acoust. Soc. Am.* 110, 2527– 2538, 2001b.

[9] K. Saberi, and D. R.Perrott. Cognitive restoration of reversed speech. *Nature*, 398, 760, 1999.

[10] B. New, C. Pallier, M. Brysbaert, L. Ferrand. Lexique 2: A New French Lexical Database *Behavior Research Methods, Instruments, & Computers,* 2005.