Available online at www.sciencedirect.com
SCIENCE DIRECT
Hearing Research 211 (2006) 74–84
HEARING RESEARCH
www.elsevier.com/locate/heares

Research paper

Masking release for consonant features in temporally fluctuating background noise

Christian Füllgrabe [a,*], Frédéric Berthommier [b], Christian Lorenzi [a,c]

[a] *Laboratoire de Psychologie Expérimentale – UMR CNRS 8581, Institut de Psychologie, Université René Descartes – Paris 5, 71 Avenue Vaillant, 92774 Boulogne-Billancourt, France*
[b] *Institut de la Communication Parlée, UPRESA CNRS 5009, INPG, 46 Avenue Viallet, 38031 Grenoble, France*
[c] *Institut Universitaire de France, France*

Received 5 September 2005; received in revised form 5 September 2005; accepted 14 September 2005
Available online 8 November 2005

Abstract

Consonant identification was measured for normal-hearing listeners using Vowel–Consonant–Vowel stimuli that were either unprocessed or spectrally degraded to force listeners to use temporal-envelope cues. Stimuli were embedded in a steady state or fluctuating noise masker and presented at a fixed signal-to-noise ratio. Fluctuations in the maskers were obtained by applying sinusoidal modulation to: (i) the amplitude of the noise (1st-order SAM masker) or (ii) the modulation depth of a 1st-order SAM noise (2nd-order SAM masker). The frequencies of the amplitude variation f_m and the depth variation f'_m were systematically varied. Consistent with previous studies, identification scores obtained with unprocessed speech were highest in an 8-Hz, 1st-order SAM masker. Reception of voicing and manner also peaked around $f_m = 8$ Hz, while the reception of place of articulation was maximal at a higher frequency ($f_m = 32$ Hz). When 2nd-order SAM maskers were used, identification scores and received information for each consonant feature were found to be independent of f'_m. They decreased progressively with increasing carrier modulation frequency f_m, and ranged between those obtained with the steady state and the 1st-order SAM maskers. Finally, the results obtained with spectrally degraded speech were similar across all types of maskers, although an 8% improvement in the reception of voicing was observed for modulated maskers with $f_m < 64$ Hz compared to the steady-state masker. These data provide additional evidence that listeners take advantage of temporal minima in fluctuating background noises, and suggest that: (i) minima of different durations are required for an optimal reception of the three consonant features and (ii) complex (i.e., 2nd-order) envelope fluctuations in background noise do not degrade speech identification by interfering with speech-envelope processing.

Keywords: Speech perception; Background noise; Masking release; Modulation masking; 1st-order modulation; 2nd-order modulation

1. Introduction

Over the last decades, many studies have compared speech identification in steady state and temporally fluctuating backgrounds presented at the same level (e.g., Miller and Licklider, 1950; Duquesnoy, 1983; Festen and Plomp, 1990; Takahashi and Bacon, 1992; Gustafsson and Arlinger, 1994; Nelson et al., 2003; Qin and Oxenham, 2003). These studies demonstrated that, in normal-hearing listeners, speech identification performance and speech reception thresholds (SRTs) were better in fluctuating than in steady-state backgrounds. At least six different mechanisms or

Abbreviations: ANOVA, analysis of variance; f_m, 1st-order modulation frequency; f'_m, 2nd-order modulation frequency; SAM, sinusoidal amplitude modulation; SD, standard deviation; SNR, signal-to-noise ratio; SRT, speech reception threshold; Tukey HSD test, Tukey Honestly Significant Difference test

* Corresponding author. Present address: Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, United Kingdom. Tel.: +44 122 376 5283.
E-mail address: c.fullgrabe@psychol.cam.ac.uk (C. Füllgrabe).

0378-5955/$ - see front matter © 2005 Elsevier B.V All rights reserved.
doi:10.1016/j.heares.2005.09.001

factors seem to be involved in this so-called "masking-release" effect.

1.1. Dip listening

Masking release is greater for square-wave than for speech-envelope modulation of a background noise (Bacon et al., 1998). It increases with the modulation depth of the modulated noise (e.g., Howard-Jones and Rosen, 1993a; Gustafsson and Arlinger, 1994) and the optimal amplitude modulation frequencies for observing masking release range from about 10 to 25 Hz (e.g., Miller and Licklider, 1950; Gustafsson and Arlinger, 1994; Kwon and Turner, 2001; Nelson et al., 2003) depending somewhat upon the speech material. Moreover, masking release increases when introducing spectral dips in a steady state or temporally fluctuating noise, and when the width of those spectral dips is increased (Peters et al., 1998). Finally, masking release is weaker or abolished for multi-talker babble maskers with a relatively flat temporal envelope, and it is reduced if the lower level portions of speech are masked by a spectrally shaped steady-state noise (e.g., Eisenberg et al., 1995; Bacon et al., 1998). These findings strongly suggest that listeners are able to take advantage of relatively short temporal minima in the fluctuating background to detect speech cues, a capacity often referred to as "listening-in-the-dips" or "listening-in-the-valleys". Clearly, this capacity requires a certain degree of temporal resolution (i.e., an ability to follow the background fluctuations in order to extract speech cues during the background valleys) and spectral resolution (i.e., an ability to access parts of the speech spectrum that are not masked or are less masked by the background). Consistent with this notion, the decrease in masking release observed for modulation frequencies greater than 30 Hz (and thus, for background valleys shorter than about 17 ms) can be attributed to forward masking – an important factor limiting temporal resolution – smoothing out or "filling in" the background valleys (e.g., Festen, 1993; Dubno et al., 2002).

1.2. Modulation masking/modulation interference

A recent speech-perception study reveals that dip listening is counterbalanced by the perceptual interference of temporal modulations in the fluctuating background with the auditory processing of temporal modulations in the speech envelope (Kwon and Turner, 2001). However, this perceptual interference does not seem to follow the general characteristics reported in previous psychoacoustical studies on modulation masking or modulation detection interference (e.g., Bacon and Grantham, 1989; Houtgast, 1989; Yost et al., 1989; Ewert and Dau, 2000; Millman et al., 2002): the modulation frequencies of speech and the fluctuating background do not allow prediction of the perceptual interference, and the amount of perceptual interference decreases as speech and fluctuating background get closer in the audio-frequency domain (Kwon and Turner, 2001). This suggests that the mechanisms supposed to be involved in modulation masking, modulation interference, and modulation tuning [e.g., modulation filters as proposed by Dau et al. (1997)] may be different from those involved in masking release. For instance, fluctuations in the background may cause some form of perceptual illusion of attentional origin leading listeners to process the background as part of the speech stimulus (e.g., Kwon and Turner, 2001; Nelson et al., 2003).

1.3. Auditory grouping

Additional studies reveal that grouping mechanisms (e.g., Bregman, 1990) may be involved in masking release since the latter disappears in case of degraded spectral resolution and loss of temporal fine-structure information (as encountered in cochlear implant users or when using noise-vocoded speech) (e.g., Nelson et al., 2003; Qin and Oxenham, 2003; Zeng et al., 2005). The fact that under these conditions speech identification is similar for fluctuating and steady-state backgrounds suggests that, when spectral resolution is sufficiently fine and temporal fine-structure information is available, normal-hearing listeners do take advantage of fine-structure cues and the incoherence of envelope cues across the spectrum to segregate speech from background.

1.4. Perceptual restoration

Earlier work demonstrated that human listeners perceive missing phonemes by using the redundancies in speech stimuli occurring at the acoustic, phonetic, phonological, and/or lexical level (e.g., Warren, 1970). It is very likely that the ability to reconstruct speech or, in other words to infer the complete spectro-temporal structure of speech from incomplete information, plays a role in this situation. For instance, Howard-Jones and Rosen (1993b) demonstrated that, to a certain extent, normal-hearing listeners are able to "patch together" information in different broad frequency regions at different times to achieve a release from masking (a process termed "uncomodulated glimpsing" by the authors). Moreover, speech identification for normal-hearing listeners is very robust for speech subjected to periodic interruption by silence or by noise segments (e.g., Miller and Licklider, 1950; Powers and Speaks, 1973; Powers and Wilcox, 1977; Nelson and Jin, 2004).

1.5. Informational masking

Recent work by Summers and Molis (2004) indicated that competing speech yields more masking than time-reversed speech containing temporal fluctuations of equal magnitude. This suggests that informational masking, presumably resulting from competitive processing of linguistic information within the speech masker, may reduce the amount of masking release [however, no difference between