Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2007) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Phonetic and lexical interferences in informational masking during speech-in-speech comprehension

Michel Hoen *, Fanny Meunier, Claire-Léonie Grataloup, Nicolas Grimault, Fabien Perrin, Xavier Perrot, François Pellegrino, Lionel Collet

Phonak AG, Audiology and Training Competence Center, Laubisruetistrasse, 28, 8712 Staefa, Switzerland

Received 2 November 2006; received in revised form 14 May 2007; accepted 19 May 2007

Abstract

This study investigates masking effects occurring during speech comprehension in the presence of concurrent speech signals. We examined the differential effects of acoustic–phonetic and lexical content of 4- to 8-talker babble (natural speech) or babble-like noise (reversed speech) on word identification. Behavioral results show a monotonic decrease in speech comprehension rates with an increasing number of simultaneous talkers in the reversed condition. Similar results are obtained with natural speech except for the 4-talker babble situations. An original signal analysis is then proposed to evaluate the spectro-temporal saturation of composite multitalker babble. Results from this analysis show a monotonic increase in spectro-temporal saturation with an increasing number of simultaneous talkers, for both natural and reversed speech. This suggests that informational masking consists of at least acoustic–phonetic masking which is fairly similar in the reversed and natural conditions and lexical masking which is present only with natural babble. Both effects depend on the number of talkers in the background babble. In particular, results confirm that lexical masking occurs only when some words in the babble are detectable, i.e. for a low number of talkers, such as 4, and diminishes with more talkers. These results suggest that different levels of linguistic information can be extracted from background babble and cause different types of linguistic competition for target-word identification. The use of this paradigm by psycholinguists could be of primary interest in detailing the various information types competing during lexical access.

Keywords: Cocktail party; Speech-in-speech; Energetic masking; Informational masking; Lexical competition

1. Introduction

Beyond the fact that we are very efficient at understanding speech delivered via headphones in the *sounds of silence* of an anechoic room, we are more usually confronted with situations where speech occurs in the acoustic chaos of a babbling crowd and yet are still able to understand the messages it transmits. The ability to segregate and understand speech despite the presence of concurrent noise or discussions is commonly referred to as the 'cocktail party' effect. Since its first description in the seminal paper by Cherry (1953), this phenomenon has given rise to an impressive number of experiments, mainly focused on psychoacoustic aspects of auditory scene analysis. These studies have provided extensive insights into the processes involved in auditory stream segregation (see for example Bregman, 1994; Divenyi, 2004a; Wood and Cowan, 1995; for reviews). In the specific context of speech-in-speech comprehension (Bronkhorst, 2000), two different but related types of masking effects must be considered, namely *energetic masking* and *informational masking* (Brungart, 2001a; Brungart et al., 2001). *Energetic masking* is attributed to the spectro-temporal composition of concurrent sounds. It occurs whenever speech is produced in the presence of a broadband noise that at least partially overlaps with it in time and frequency. *Informational masking* is related to the type of information present in interfering

* Corresponding author. Tel.: +41 44 928 07 96; fax: +41 44 928 07 07.
E-mail address: ~~Michel.Hoen@phonak.com~~ (M. Hoen).

0167-6393/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.specom.2007.05.008

sounds (Dirks and Bower, 1969; Festen and Plomp, 1990). Although there is not necessarily any physical overlap in the signals from target- and masker-sounds, a competitive aspect is introduced during the later processing of these signals. In the context of speech-in-speech comprehension, some energetic masking certainly does occur, although this has recently been shown to be responsible for only a relatively small part of the overall masking phenomenon which occurs in this listening condition (Brungart et al., 2006). This highlights the importance that informational masking of concurrent speech signals can have on the intelligibility of target speech signals. Moreover, it appears that energetic masking is even less prominent during speech-in-speech comprehension than when speech is presented together with speech-modulated noise (Brungart et al., 2006). This observation suggests that informational masking plays a predominant role during speech-in-speech comprehension and stresses the need for more extensive study of the relevant processes.

The intelligibility of one target speech signal presented diotically (the same signal is presented to both ears) and in a background of other speech signals is modulated by two co-varying factors: (1) the number of simultaneous talkers and (2) the temporal envelope of the resulting babble (Bronkhorst, 2000). When only a few simultaneous talkers are present, listeners can take advantage of asynchronies in the dynamic variations of the different concurrent streams causing transient gaps in the babble during which they can listen to target-signals. However, as the number of talker increases, the dynamic modulations present in the additive sources are progressively averaged. The duration of gaps in the composite babble thus decreases, leading to a progressive shrinking of the temporal window available for listening to the target signal (e.g. Bronkhorst and Plomp, 1992; Drullman and Bronkhorst, 2000; Hawley et al., 1999; Miller, 1947; Peissig and Kollmeier, 1997). Ultimately, adding more talkers will lead to speech-modulated noise. It is obvious that the most complex informational masking phenomena must happen at a relatively low number of background-talkers, while some speech-specific information is still available from babble sounds and can compete with target information. Up to now, only a few studies have investigated informational masking in detail. In particular, the effect of additional multiple masker-talkers in the concurrent background on the amount and type of informational masking effect that occurs during speech-in-speech comprehension is still only partially known. This study tested the assumption that informational masking results from both acoustic–phonetic and lexical masking effects whose magnitudes depend on the number of talkers in the concurrent noise. To assess this hypothesis, a word comprehension task was proposed, in the presence of several concurrent noises: 4- to 8-talker babble (natural speech) and babble-like noise (reversed speech) yielding various amounts of acoustic–phonetic and lexical information. In the second section of the present introduction we report recent results on the informational masking phenomenon and present the current study. Section 3 is a thorough description of the data and experimental design of the behavioral experiment. Section 4 describes the results from the behavioral study investigating the influence of lexical and phonemic information present in background babble during the comprehension of single words against multitalker babble sounds. An original measure for spectro-temporal saturation is then introduced. It was specifically designed to analyze the effect of increasing the number of talkers in natural and reversed babble. Section 5 discusses the results of this study in the light of the initial assumptions.

2. Informational masking during speech-in-speech comprehension

2.1. Recent results

Brungart (2001a) and Brungart et al. (2001), started addressing the issues of relative intensity of the different talkers, talker number, talker gender, and signal-to-noise ratio (SNR) in situations with up to three background-talkers. In their experiments, all stimuli were short sentences extracted from a corpus first developed by Bolia et al. (2000) and built on the general structure: 'Ready (*call-sign*)? Go to (*color*) (*number*) now.'. The complete set of phrases contains all combinations of 8 different call-signs ('Arrow', 'Baron' or 'Charlie' for example), 4 colors ('red', 'green', 'white' and 'blue') and 8 numbers (1–8). A typical sentence in this corpus would be: '*Ready Baron? Go to green eight now*.' Participants were informed that the target signal was the one starting with '*Ready Baron?*' whereas the concurrent stimuli always used one of the other 7 call-signs. The listener's task was to select on a computer screen the colored digit corresponding to the color and number used in the target sentence (Brungart, 2001b). Results from their first study, using a listening situation with 1 target-talker against 1 masker-voice, evidenced the main contribution of informational masking in this situation; listeners were generally able to hear both competing speech messages, but experienced difficulty separating the content of the target phrase from that of the masking phrase (Brungart, 2001a). This study also showed that listeners could use differences in the intensity levels of the two talkers to selectively hone in on the quieter voice stimulus, and that consequently SNR had relatively little influence on the intelligibility of the target talker for SNRs in the 0 dB to −10 dB range (see also Dirks and Bower, 1969; Egan et al., 1954). Brungart et al. (2001) extended their initial observations to experiments with 2-talker and 3-talker backgrounds. Results showed a global linear decrease in performance with decreasing SNR and a strong effect of talkers' gender; performances were worst when the target talker and the babble were of identical gender, in particular at positive SNRs. These experiments all show that in the presence of a low number of background-talkers (1–3), masking effects caused by concurrent voices are easily overcome, certainly based on clear acoustical distinctions such as differential intensity or pitch. In these conditions, the most important information used to compensate masking phenomena are vocal characteristics such as relative intensity or gender type. It is generally acknowledged that certain vocal characteristics such as gender are encoded in pitch information. It therefore appears that pitch similarities are responsible for the first main component of informational masking and that conversely pitch differences are used to segregate speech streams when the number of masker-talkers does not exceed 3 (see also Divenyi, 2004b). Similarly in their study on the processing of stressed and unstressed syllables against speech-modulated noise, Divenyi and Brandmeyer (2003) showed that stressed syllables constituting local improvements in SNR were better recognized than unstressed syllables, reinforcing the idea that prosodic cues were highly relevant in this context. The above mentioned studies all highlight the first level of the informational masking effects which occur during speech-in-speech comprehension, namely prosodic masking. If target- and masker-voices have relatively close pitch, they are even harder to separate and so identification of the target-signal is even harder. However, besides pitch, other psycholinguistic dimensions of speech sounds may well play an important part in informational masking. The existence of phonemic or lexical interference in multitalker speech comprehension situations implicating high numbers of concurrent voices for example, compared to those observed for other noise backgrounds has to our knowledge rarely been considered. In a recent study however, Simpson and Cooke (2005) measured consonant identification rates in a closed set of vowel–consonant–vowel tokens gated with multitalker babble noise and babble-modulated noise. In this experiment, authors used babble noise made up of 1–512 talkers. Their results showed a non-monotonic decrease in performance with increasing numbers of talkers. Globally, for babble noise, performances first decreased gradually with increasing numbers of talkers, reaching minimal values for 6 talkers. The results for babble noise made up of 8–128 talkers then remained almost stable, the resulting masking effect remaining approximately the same in all these conditions. On the contrary for babble-modulated noise, performances fell gradually. In this experiment a maximal difference between masking effects caused by the babble-modulated noise and natural babble was observed for 8 talkers where noise was associated with approximately 18% better recognition scores than natural babble. This experiment clearly suggests that increased informational masking occurs, certainly attributable to relevant acoustic–phonetic information present in the natural babble conditions competing with target CVC information. This competition causes a worsening in performance that is a non-monotonic function of the number of talkers in the background, starting when at least 3-talkers are present in the background and reaching maximal levels when 8 talkers compose the background. However, one limitation of the work proposed by Simpson and Cooke is that they did not use real words, which prevented them from drawing conclusions from their data on the processes implicated in lexical activation against background babble sounds. The current work therefore aimed to characterize the interferences produced by different types of speech or speech-derived backgrounds on word identification performance. In particular, we wanted to identify if and how lexical and acoustic–phonetic information specifically interferes during multitalker speech-in-speech comprehension and how this interference is modulated by the number of talkers. The nature of informational masking in the particular case of speech-in-speech comprehension is still unspecified, mainly because the specificities of linguistic information have rarely been considered in this context. For example, the experiments of Brungart (2001a) and Brungart et al. (2001) used a paradigm that from a psycholinguistic viewpoint was rather minimalist, as the experimental linguistic material was based on a small predetermined closed set of sentences repeated throughout the experiment. Moreover, the speech comprehension task consisted rather in a choice within a closed set of propositions than identification *per se*. In our experiment, we decided to take the psycholinguistic parameters of target words into account (see Section 2.2) and to use natural diversified speech signals to create babble sounds.

2.2. The present study

Our experiment studied the nature and availability of information present in a babble background and interference with word identification when the number of simultaneous talkers was increased. As Brungart et al. (2001) showed that with up to 3-talkers, masking is mostly due to the processing of pitch information that possibly interacts with or even covers the masking effects due to other psycholinguistic dimensions of speech sounds, we decided to study the cases where individual voice characteristics are less predominant, i.e. situations with 4 or more talkers. We contrasted situations where the babble was made of natural speech and therefore contained real words (Natural Speech) vs. situations in which only partial phonetic information was available (reversed speech) vs. situations in which no phonetic information was available (Noise). As babble sounds, we used signals composed of 4, 6 and 8 simultaneous talkers (S4, S6 and S8) of mixed gender (50% male, 50% female voices). To control for potential effects of target and mask gender information we also included one same (masculine) gender condition in the case of 4-simultaneous talkers (S4m). In order to dissociate the spectro-temporal saturation effect from a potential lexical masking effect we also took advantage of using the same speech sounds but reversed along their temporal axis (reversed babble sounds, later referred to as R4, R6, R8 and R4m). Time reversal of speech signals has been claimed to be the most drastic degradation one can apply to speech (Saberi and Perrott, 1999). However, not only

Please cite this article in press as: Hoen, M. et al., Phonetic and lexical interferences in informational masking ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.05.008

does reversed speech 'sound' like speech, it also keeps partial phonetic information present in natural speech (like vowels or fricatives for example). Moreover, when different reversed speech streams are mixed together, the resulting babble sounds like normal speech babble and one can clearly perceive phonemes in it, although it does not contain actual words. Reversed babble stimuli were thus considered in the experiment as control speech sounds containing phonetic but no lexical information. To further obtain a reference measure of a pure energetic masking effect and dissociate it from the different levels of informational masking potentially identifiable in this experiment, we added one condition where speech was presented against a broadband noise background (later referred to as N). This noise was made to have spectro-temporal characteristics similar to our most spectro-temporally saturated natural and reversed babble signals (e.g. S8 and R8). These 9 background noise types (S4, S6, S8, S4m, R4, R6, R8, R4m and N) were all tested at 4 different SNRs of −3, 0, +3 and +6 dB, yielding a total of 36 main experimental conditions.

3. Materials and methods

3.1. Concurrent sounds (Fig. 1)

3.1.1. Multitalker babble sounds and reversed babble sounds

Mixed groups of 50/50 male/female talkers were created for 4, 6 and 8 voices and one group of 4 males; these groups gave the babble signals. Each voice was first recorded separately in a sound-proof room, reading extracts from the French press. Individual recordings were modified according to the following protocol: (i) removal of silences and pauses of more than 1 s, (ii) suppression of sentences containing pronunciation errors, exaggerated prosody or proper nouns, (iii) noise reduction optimized for speech signals, (iv) intensity calibration in dB-A and normalization of each source at 80 dB-A and (v) final mixing of individual sources into cocktail party sound tracks. Reversed babble sounds were obtained by reversing the previously generated speech babble stimuli along their temporal dimension.

3.1.2. Associated broadband noise

In order to obtain a broadband noise with spectro-temporal characteristics comparable to those of our most saturated natural and reversed babble, we used the 8-talker babble as reference. This signal exhibited the strongest energetic masking effect because it had both the richest spectral composition of all our babble sounds and contained the highest number of independent auditory streams. Envelope information from the original speech babble was extracted below 60 Hz. Using Fast Fourier Transform (FFT), the power spectrum and phase distributions 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 babble. Finally, the root mean square (rms) powers of the original and new signals were matched.

3.2. Target words

Two hundred and eighty-eight French mono-syllabic, tri-phonemic words were recorded in a sound-proof booth by a male native French speaker. Words were selected in a middle range of frequency of occurrence (ranging from 0.19 to 146.71 per million; mean = 20.96, SD = 21.37), according to the French database Lexique2 (New et al., 2004), in order to avoid extremely high- or low-frequency items that volunteers typically either overuse or ignore.

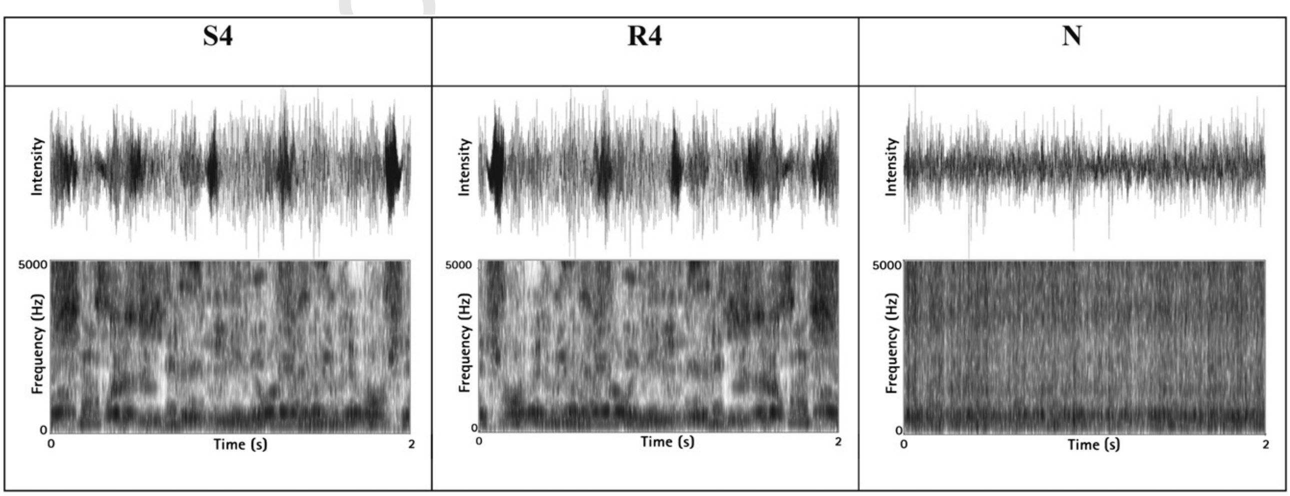

Fig. 1. Waveform and spectrogram of three examples of background noises used in the behavioral experiment. From left to right: 4-talkers natural babble (S4); 4-talkers reversed babble (R4) and broadband speech-modulated noise (N).

3.3. Stimuli and word lists

Stimuli consisted of the 288 single target words mixed together with 4 s of background noise. Target words were always inserted 2.5 s from the start of the stimulus, so that participants always had the same exposure to the background noise before the target word was presented. Stimuli were composed by mixing one chunk of background noise, randomly selected from 40 chunks extracted from the original noise files, with one target word. Individual intensity levels for background noise and target words were adjusted according to the global rms power of the original sounds to be mixed. As this resulted in some modulation of the intensity of the final stimulus and in order to avoid the global intensity of stimuli becoming predictive of the SNR, a final randomized intensity roving over a $+/-3$ dB range in 1 dB steps was applied to every created stimulus.

Thirty-six different lists – one for each participant – were generated, each list containing every target word only once to avoid priming effects. Across lists, all target words were presented against the 36 background conditions. Finally, each list was made up of 288 stimuli, 8 in each of the 36 experimental conditions. Within lists, target words were balanced for frequency and phonological neighborhood across conditions.

3.4. Participants and procedure

Thirty-six volunteers participated in the experiment, they were all students, aged 18–32 years and native French speakers with no known hearing or language disorders. They were paid for their participation. Participants sat in a quiet room, facing a computer monitor. Stimuli were delivered diotically via headphones (Beyerdynamic DT 48, 200 Ω) at an individually adjusted comfortable sound level. The task for participants consisted in a single-word transcription, participants being asked to type the sounds they heard on a computer keyboard. Before the testing phase, participants were given 12 practice items to accommodate themselves to the stimulus presentation mode and target's voice. The experiment lasted an average 45 min.

4. Results

4.1. Speech-in-speech comprehension: behavioral results

Answers from participants were analyzed in terms of correct word identification rates by calculating the proportion of transcribed words that corresponded to target words. These individual word identification rates were used as dependant variables in the following analyses.

For all the analyses, it appeared that the gender factor (masculine, mixed gender) was never significant either as a main effect, or interaction (all $p > .05$). As an example, when directly comparing 4-talker babble and reversed babble and including the mixed and same gender conditions, we observed a main effect of babble conditions ($F(1,35) = 30.106$, $p < .0001$), but no effect of gender ($F(1,35) = 1.191$, n.s.) and no interaction ($F < 1$). We have therefore not included the gender factor in the following analyses.

As a first step and in order to compare the effect of the three types of background noise (broadband noise, natural and reversed babble), we ran a within-subject repeated measures ANOVA, taking into account performances observed for the broadband noise condition and the natural and reversed 8-talker babble conditions. We chose to compare the broadband noise condition to the 8-talker babble conditions as the noise was an 8 babble-modulated broadband noise (see Section 3.1.2). Descriptively, we observed that word identification was better in broadband noise (69%, SD = 15) than in reversed or natural babble (see Fig. 2) which gave similar results (59% and SDs = 14 and 16) respectively. We observed a significant main effect of the type of background noises ($F(2,70) = 20.285$, $p < .0001$) and a significant main effect of SNRs ($F(3,105) = 134.831$, $p < .0001$). There was a monotonic increase in identification performance with increasing SNR. The interaction between these two factors was not significant ($F(6, 210) = 1.405$, n.s).

We next compared natural and reversed babble more directly (S4, S6, S8 vs. R4, R6 and R8). In this 3-way within-subject repeated measures ANOVA, we included as factors: Babble type (natural and reversed), Number of talkers (4, 6 and 8) and SNR (-3, 0, 3 and 6). This analysis revealed a significant main effect only for the SNR factor ($F(3,105) = 185.309$; $p < .0001$), word identification rates globally decreasing monotonically with the SNR in our range. Main effects for the factors Babble type ($F(1,35) = 1.176$; n.s.) and Number of talkers ($F(2,70) = 2.190$; n.s.) remained non-significant, but the second level interaction between these two factors was significant ($F(2,70) = 4.534$; $p < .02$), revealing that the effect of reversing the speech signals along their temporal dimension interacted with the effect of increasing numbers of talkers present in the babble. All other interactions remained non-significant; in particular, the SNR factor did not interact with any other factor.

Planned comparisons performed on the second level interaction between the factors Babble type and Number of talkers revealed that word identification rates observed in the natural 4-talker babble condition were significantly lower than those observed in the reversed 4-talker babble ($F(1,35) = 9.469$, $p < .005$) condition while the other differences between natural and reversed babble remained non-significant (S6 vs. R6: $F(1,35) = 1.137$, n.s. and S8 vs. R8: $F(1,35) < 1$). Indeed the S4 and R4 conditions had very different impacts on average word identification rates, S4 being the condition associated with the poorest word identification performances amongst natural babble sounds (57% vs. 63% for S6 and 59% for S8) and conversely R4 the condition associated with the best performances amongst reversed babble (63% vs. 61% for R6 and 59% for R8) (see Fig. 3). Actually, S4 was the background noise

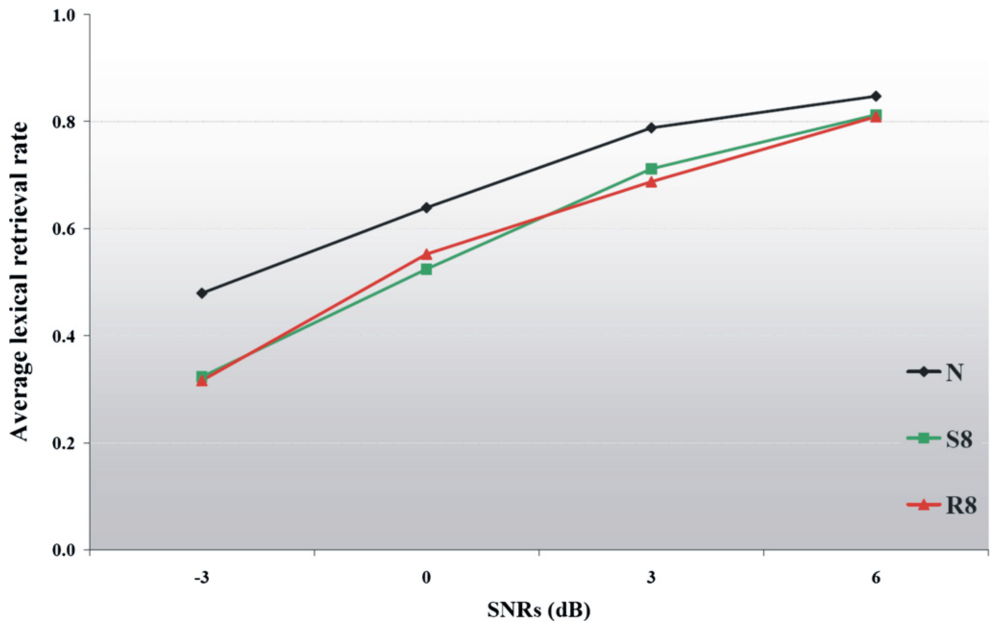

Fig. 2. Word identification rates for the broadband noise condition (condition N) and the natural and reversed 8 talker babble (conditions S8 and R8) depending on SNR.

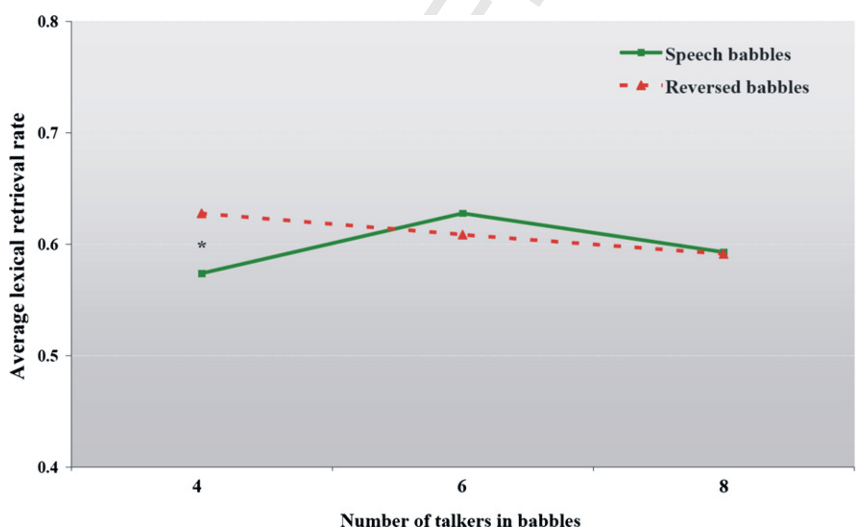

Fig. 3. Direct comparison of informational masking effects due to increasing number of voices present in natural and time-reversed babble (conditions S4, S6, S8, R4, R6 and R8). Note in particular the difference between the natural and reversed 4-talker babble (S4 and R4).

which caused the greatest masking effect while R4 caused the least for *all* babble and babble-like conditions. This demonstrated that S4 and R4 did not mask speech items with equal efficiency although their spectro-temporal structures were very similar. It is also worth noticing that for reversed babble, R4 allowed better word identification than did R6 which in turn was better than R8 while for natural babble, S4 and S8 gave comparable lower performances than did S6.

Since the most important informational masking effect is obtained with lower numbers of talkers, we argue that in the 4-talker babble, more information or information of a particular nature is available from the babble and that this competes during the target word identification processes, causing an increased informational masking effect leading to lower word comprehension rates than in the 6-talker babble. In order to test this proposition and to further specify the nature of the interference, we performed a partial analysis of transcription errors in the responses given.

Most erroneous answers were phonological neighbors, consisting in words sharing at least one phonological unit with the target word, such as *môme* /məm/ "kid" for *paume*/pəm/ "palm" or part of the target such as *heaume* /əm/ "helmet". However, very few answers had no overlap at all with the target such as *avion* (aircraft) instead of *paume* (palm). We analyzed this type of error as it may reveal more clearly the influence of the background babble. Overall, we observed 182 errors of this particular type, representing 1.76% of all responses given in the experiment. The distribution of these errors as a function of the type of background noise and number of talkers is shown in Fig. 4.

Overall, 66.5% of these errors were observed when the background babble was composed of natural speech (S4m, S4, S6 and S8) and 29.7% in reversed speech conditions (R4m, R4, R6 and R8). The speech-in-broadband noise condition (N) reached 3.8%, giving an estimation of the proportion of errors of this type made in the absence of any speech or speech-derived background. We ran a 2-way ANOVA considering the number of words given by participants that did not overlap phonologically with the targets as dependent variables. We included as within factors Babble type (natural and reversed) and Number of talkers (4, 6 and 8) and observed an effect of the Babble type ($F(1,35) = 12.169$, $p = .001$), an effect of the Number of talkers ($F(2,70) = 3.818$, $p < .05$) but also an interaction between the two factors ($F(2,70) = 3.855$, $p < .05$). Specific comparisons revealed a significant difference between S4 and R4 ($F(1,35) = 16.579$, $p < .001$); more words phonologically unrelated to the target word were proposed in S4. No difference was observed between S6 and R6 ($F(1,35) = 2.015$, n.s.) and between S8 and R8 ($F(1,35) = 1.044$, n.s.). The difference observed between S4 and R4 showed that more words, phonologically unrelated to the targets, were produced in the natural babble condition than in the reversed one. Moreover, looking at the words proposed in S4 it appears that 52.63% ($n = 20$) of them belonged to the babble.

4.2. Preliminary discussion of the behavioral results

It seems that different major effects of background sounds can be identified in this complex pattern of results. It appears that the first predominant factor influencing speech comprehension in multitalker situations is of a quantitative type and consists of a spectro-temporal saturation effect of the babble that causes progressively increasing masker effects due to a decrease in the temporal gaps free for listening to target words. This phenomenon could be considered as informational masking occurring at the acoustic–phonetic level; it would explain the monotonic worsening of scores when the number of talkers in reversed babble signals increases (63% for R4 vs. 61% for R6 and 59% for R8) and similar behavior would be expected for normal speech babble (S4 to S8), although the significant differences observed between R4 and S4 suggest that a second factor is operating. We will assume that this factor is a matter of lexical competition initiated by words detectable in the S4 condition as our partial analysis of errors clearly demonstrated the activation of words from the babble that compete as lexical candidates with the target-word during identification processes.

To evaluate whether this hypothesis of a similar acoustic–phonetic masking for S and R conditions was correct, we proposed a method designed to precisely evaluate the effect of spectro-temporal saturation caused by increasing the number of talkers present in a multitalker babble sound. We therefore ran an acoustic analysis to highlight this effect and assess its potential influence on speech comprehension.

4.3. Acoustic analysis of natural and reversed babble

The general shape of a speech signal is characterized, among other things, by its spectral envelope and its associated temporal fluctuations (e.g. Greenberg, 1995). Yet as

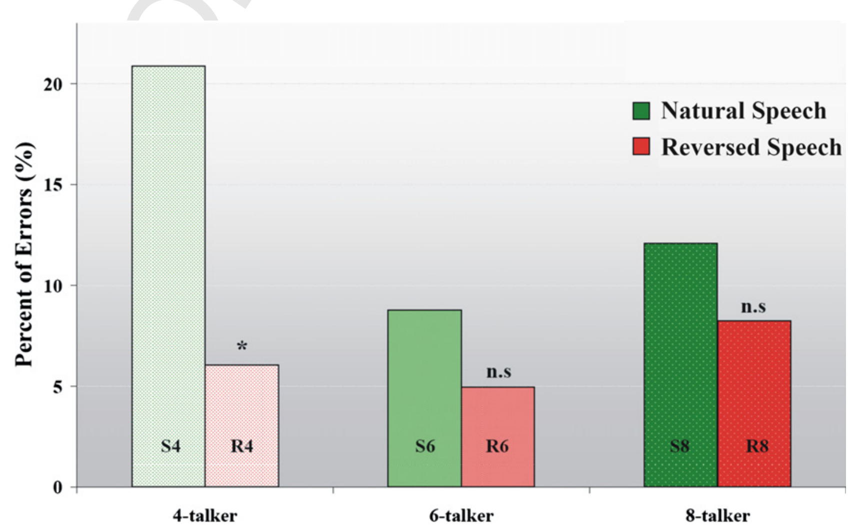

Fig. 4. Distribution of errors consisting in words not sharing any phonemic characteristics with the target word as a function of the type of background noise and number of speakers.

soon as several talkers overlap, this leads to progressive spectro-temporal saturation and a reduction in both the amplitude and period of the fluctuations. Evaluating the intrinsic properties of this kind of multitalker signal is not straightforward. Previous work has popularized the measurement of speech intelligibility through the evaluation of temporal modulations, often referring to the notions of modulation spectrum, envelope spectrum or modulation transfer function (Chi et al., 1999; Greenberg and Arai, 2001; Houtgast and Steeneken, 1985; Steeneken and Houtgast, 1980). Broadly speaking, speech signals exhibit peak sensitivity between 3 Hz and 4 Hz, and this specific attribute has been used for a wide range of purposes, from the evaluation of intelligibility through a transmission channel (Houtgast and Steeneken, 1985) to discrimination between speech and music (Karnebäck, 2001; Pinquier et al., 2003; Scheirer and Slaney, 1997). These approaches find their roots in both the auditory bases of speech perception (Bacon and Viemeister, 1985) and the study of syllabic duration in different languages (Greenberg et al., 1996). Although the impact of the transmission channel (additive noise, reverberation) on this envelope spectrum has been extensively studied, the variability of the modulation spectrum for both speech and multitalker babble *per se* remains unknown. Consequently, these measurements and their extensions like the STI (Speech Transmission Index) propose no direct way to analyze the intrinsic properties of multitalker babble but only an indirect way to assess their impact on speech-in-speech intelligibility.

As an alternative to performing an intrinsic analysis on these babble signals, we propose an original approach directly inspired by the acoustic–phonetic analyses used for automatic speech recognition. Its purpose is twofold: (1) to identify the size of the "units" of temporal coherence in the waveform and (2) to evaluate the spectral relationship existing within a sequence of such "units". Our hypothesis is that multitalker babble will pattern differently from single-speaker speech along these spectro-temporal dimensions. More specifically, multitalker babble should not exhibit numerous regularity patterns as does speech in terms of temporal structure: as soon as several streams overlap, the typical alternation of well-formed phonemes should be degraded and the more streams are merged, the more degradation should be observed.

4.3.1. Methods

In order to estimate the temporal structure of the different signals, a statistical segmentation method was applied, namely the 'Forward–Backward Divergence' algorithm (André-Obrecht, 1988). This method is based on the measurement of the Kullback–Lieber divergence between two autoregressive models evaluated on two different but overlapping temporal windows and is designed to detect abrupt changes in waveforms. When applied to a single-speaker signal, it identifies boundaries strongly related to the phonetic structure of speech and defines two main categories of sub-phonemic segments: short segments (bursts and transient parts of voiced or unvoiced phonemes) and long segments (steady parts of phonemes). Comparing the size of these segments in different types of babble (in terms of number of speakers) may consequently give us information on the temporal coherence of signals. Interested readers are referred to André-Obrecht's (1988) paper for a comprehensive and detailed description of this algorithm.

Besides the fact that increasing the number of talkers in babble sound should blur its temporal structure by mixing asynchronous phonetic streams, it should also result in the smoothing of spectral differences between successive segments, as identified by the segmentation algorithm. In order to assess this effect, we performed a cepstral analysis in 8 Mel frequency filters in the middle of each segment. A Euclidean distance was then computed in the Mel Frequency Cepstral Coefficient (MFCC) space between consecutive segments, providing a measure of the acoustic–phonetic coherence of the signal.

4.3.2. Materials and parameters

Analyses were performed on mixtures of 1, 2, 4, 6 and 8 talkers and their time-reversed equivalents (10 conditions). Each sound track was generated according to the mixing procedure employed to generate babble stimuli used in the behavioral experiment. Original sounds were then normalized at a level of −20 dBFS and split into 5 s chunks. Sixty chunks were randomly selected for each condition and two parameters $\overline{D}_{\text{seg}}(c)$ and $\overline{\Delta}_{\text{cep}}(c)$ were extracted for each chunk c:

$$\overline{D}_{\text{seg}}(c) = \frac{D(c)}{N_{\text{seg}}(c)} \tag{1}$$

where $D(c)$ is the duration (in ms) and $N_{\text{seg}}(c)$ the number of segments determined by the segmentation algorithm for chunk c. $\overline{D}_{\text{seg}}(c)$ is therefore the average segmental duration within the chunk.

$$\overline{\Delta}_{\text{cep}}(c) = \frac{1}{N_{\text{seg}}(c) - 1} \times \sum_{k=2}^{N_{\text{seg}}(c)} d_{\text{cep}}(s_{k-1}, s_k) \tag{2}$$

where $s_i(c)$ (simplified to s_i for clarity) is the ith segment of chunk c and $d_{\text{cep}}(.,.)$ is the Euclidean distance computed in the multi-dimensional MFCC space. $\overline{\Delta}_{\text{cep}}(c)$ is consequently a measure of the cepstral distance between two consecutive segments, averaged on chunk c.

Before developing these indices in the context of babble sounds, it may be useful to clarify their nature in a 1-talker situation. Let us consider a given text uttered several times by a given speaker at different speaking rates (slow vs. normal vs. fast) and different speaking styles (hyper-speech vs. normal vs. hypo-speech, see Lindblom, 1990). All other parameters being equal, changing from slow to fast rates will diminish the average length of the segments, shifting to low $\overline{D}_{\text{seg}}(c)$ values. Conversely, changing from hypo-speech to hyper-speech will increase the distance between to consecutive phonemic targets and thus shift $\overline{\Delta}_{\text{cep}}(c)$ to

Please cite this article in press as: Hoen, M. et al., Phonetic and lexical interferences in informational masking ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.05.008

higher values. In this over-simplistic scheme and for 1-talker speech, these two parameters may be associated with the speaking rate and style. However, reality is more complex and the two dimensions may interact in a complex way. Noteworthy are the two opposing predictions in the literature on the influence of speech rate on the spectral (or cepstral) distance between successive segments, depending on whether target undershooting is observed or faster movements are performed (for a review, see Wrede, 2002, pp. 27–51). To the best of our knowledge, this kind of coherence measurement has not yet been used for babble analysis. It may nevertheless be assumed that the number of streams present in the babble will influence the two parameters.

In order to subsume these parameters, a composite index called the cepstral variation rate (CVR) is then defined as

$$\mathrm{CVR}(c) = \frac{\overline{\Delta}_{\mathrm{cep}}(c)}{\overline{D}_{\mathrm{seg}}(c)} \quad (3)$$

CVR subsumes both spectral and temporal dimensions and provides a convenient index for noise structure comparison. Along this dimension, single-talker chunks will provide high CVR values, due to a relatively high numerator (possible sequence of very distant phonemes as unvoiced closures and vowels) and a relatively low denominator (sub-phonemic segmental duration). On the contrary, as the number of simultaneous talkers in babble increases, the CVR value should decrease since temporal blurring results in an increase of $\overline{D}_{\mathrm{seg}}(c)$ due to the decreasing number of abrupt changes in the signal while spectral smoothing tends to diminish $\overline{\Delta}_{\mathrm{cep}}(c)$. Intermediate values between these extreme states may occur for 1-talker speech depending on the speaking rate and style. A slow hyper-speech will result in both high segment duration and cepstral distance while a fast hypo-speech will result in low values for both variables. Consequently, both situations yield similar intermediate CVR values. In multitalker situations, a thorough study of this statement may be necessary, but it is beyond the scope of the present study since these parameters (speaking rate and style) are not manipulated in the babble.

4.3.3. Results and discussion of acoustic analysis

A 2-way between-items ANOVA considering chunks as random variables and Cepstral Variation Rate (CVR) as dependent variables was performed. We included as factors, babble Type (normal and reversed) and Number of talkers (1, 2, 4, 6 and 8). This analysis revealed an absence of significant main effect of Babble type ($F(1,589) < 1$), CVR of the considered sounds being independent of time direction. The main effect of Number of talkers was significant ($F(4,589) = 2718.691$; $p < .0001$), the CVR decreasing with the number of talkers present in the natural or reversed babble as shown in Fig. 5. The second level interaction remained non-significant ($F(4,589) < 1$) suggesting that the decrease in spectral complexity with the number of talkers was independent of the time direction of the acoustic signal.

This acoustic analysis clearly showed that an increase in the number of talkers composing babble sounds causes a proportional increase in spectro-temporal saturation as evaluated by an automatic segmentation algorithm and the evaluation of cepstral variation amongst segments. According to previous observations (Bronkhorst and Plomp, 1992; Miller, 1947) the release from masking due to *listening-in-the-gaps* made possible by dynamic fluctuations in babble sounds disappears for four or more simultaneous talkers. Therefore, a monotonic increase in spectro-temporal complexity and saturation in babble

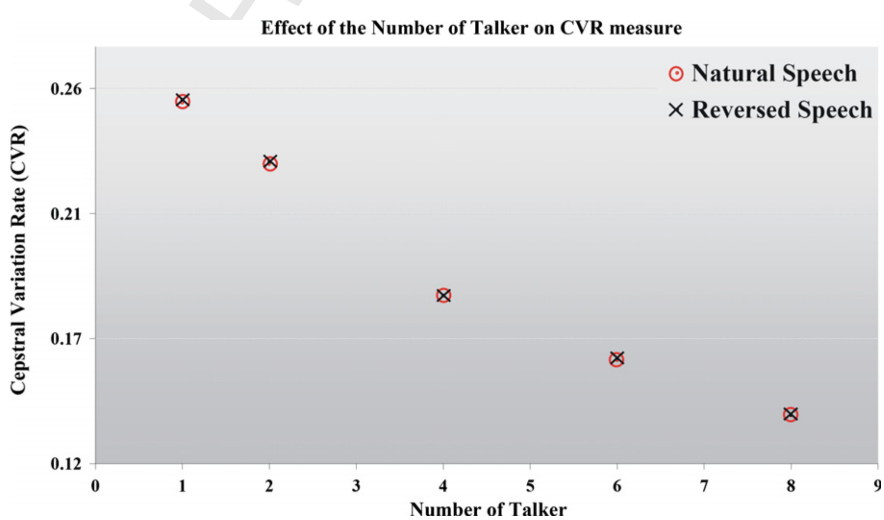

Fig. 5. Cepstral variation rate (CVR), reflecting the spectro-temporal saturation of sound for the natural (red circles) and reversed (black crosses) babble conditions, plotted against the number of talker present in babble. Observed monotonic decrease in CVR corresponds to a monotonic increase in spectral saturation with increasing number of voices. Error bars represent +/−1 SD. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

should be associated with a monotonic increase in the masking effect of that sound and consequently speech comprehension against that sound should decrease monotonically. This is what we observed in background noise conditions using reversed speech sounds (R4, R6 and R8); however, using natural speech sounds (S4, S6 and S8), the pattern was non-monotonic, R4 and R8 showing comparable masking effects and S6 globally constituting a weaker masking situation.

5. General discussion and conclusions

In this paper we were interested in speech-in-speech comprehension and more particularly in determining the influence of a background of babble on target-word identification. Behavioral results showed that globally the SNR plays a constant role over conditions; identification performance increased monotonically with SNRs. This effect did not interact with any other factor in our experiment. Concerning the masking effect produced by the different types of background noise we observed that broadband noise allows the best performance compared to matched natural and reversed babble conditions. This result suggests that linguistic interference produced by natural and reversed babble reduces the identification performance by an average of 10–20% compared to a situation where no linguistic units are involved. This result is in line with former studies that demonstrated that babble was a more effective masker than speech-shaped noise (Brungart et al., 2006; Danhauer and Leppler, 1979; Duquesnoy, 1983; Festen and Plomp, 1990; Simpson and Cooke, 2005).

Looking at the effect of the number of talkers and comparing the natural and reversed babble, we observed that while performances associated with reversed babble showed a monotonic decrease matching the increase in the number of talkers speaking, this was not the case for natural babble conditions. More specifically, significant differences were observed between the two 4-talker conditions. We designed an original acoustic analysis method that confirmed that increased spectro-temporal saturation was a monotonic effect of increasing the number of talkers in babble sounds. In diotic multitalker speech-in-speech situations, the masking effect of natural babble interferes with single-word comprehension, yet does not appear to increase monotonically with the number of talkers, at least in the range tested of 4–8. In fact, listeners recognized target words more easily from among a 6-talker babble than a 4-talker babble. Such findings are borne out by Simpson and Cooke (2005) who also observed that in the natural babble condition performances were non-monotonic, contrary to that observed in babble-modulated noise. For natural babble, reported performances rapidly fell to a minimum for 8-talker babble. Little improvement was then observed between 8- and 128-talker babble, the point at which both the natural babble condition and babble-modulated noise produced the same masking effect. Their results showed that babble-modulated noise is a less effective masker when there are at least three simultaneous talkers in the babble and beyond. In babble noise, performance decreased gradually with an increasing number of talkers. In the 8-talker babble condition comparable to the one we used, they reported about an 18% reduction in performance compared to a babble-modulated noise condition. Simpson and Cooke (2005) proposed that their results represent the combined contribution of multiple acoustic, linguistic and attention related factors to the perception of consonants. Here, our interpretation can be much more focused as the range of variation is considerably smaller.

In our experiment, the increased difficulty (around 8% of word comprehension reduction) observed for the natural 4-talker babble compared to the natural 6-talker babble is contradictory to the simple effect of progressive dynamic envelope saturation with an increasing number of simultaneous talkers in the babble. We argue that this effect is due to energetic and informational masking in speech-in-speech situations which vary with the number of talkers present in the natural babble. While energetic masking globally increases monotonically with increasing numbers of talkers, our results suggest that different types of informational masking occur depending on the number of simultaneous talkers. In order to further specify the processes implicated in these interferences, we ran a partial analysis of the word identification errors made by participants. This analysis clearly evidenced that in the 4-talker condition, significantly more words from the background babble were activated and competed with the identification of target words to be eventually given as answers. We therefore suggest that in this particular condition, the increased informational masking effect is attributable to increased lexical competition effects triggered by the availability of identifiable lexical items from background babble. With a 6-talker situation, the saturation of the background babble would then be such that complete lexical items would no longer be available or be available only to a lesser extent and would therefore cause a decrease in the global informational masking effect, causing a small improvement in performances in this condition. In Simpson and Cooke's experiment, the authors reported a decrease in performances down to the situations with 6 and 8 talkers and a relative plateau of bad performances for further increases in talkers up to 128. This difference can be accounted for by the fact that we employed real words and not CVC items. Our results suggest that with words, another layer of potential informational masking is added, constituted of lexical interferences that show a quite important effect. Using only CVC items can mainly cause acoustic–phonetic interferences to occur, but no added lexical masking effect.

From a psycholinguistic point of view, most models of lexical access, although making different proposals regarding the nature of the competitors, postulate that word identification is the result of strong competitive mechanisms between lexical candidates activated simultaneously (see for example the neighborhood activation model of Luce and Pisoni (1998) and Luce et al. (1990); the revised Cohort model of Marslen-Wilson (1987), Marslen-Wilson (1990) and Marslen-Wilson et al. (1996), and connectionist models such as TRACE proposed by McClelland and Elman (1986) and Shortlist, Norris (1994)). The effects of lexical competition have been shown in a number of studies. For instance, the number and frequency of phonetic neighbors have been demonstrated to influence the speed of response in lexical decision and shadowing, as well as influencing the percentage of correct identification in noise (Luce and Pisoni, 1998; Luce et al., 1990). Priming experiments showed reduced lexical activations through reduced facilitation effects, when prime stimuli remained ambiguous between multiple lexical candidates (Gaskell and Marslen-Wilson, 1997; Marslen-Wilson et al., 1996; Moss et al., 1997; Zwitserlood and Schriefers, 1995). In auditory lexical decisions, previous presentations of words, such as *bruise*, slowed subsequent responses to competitors, such as *broom* (Monsell and Hirsh, 1998) showing that the increased activation of a phonologically similar item could delay recognition of a competitor. These experiments have all demonstrated that successful lexical access is the result of a competition between different lexical candidates (see also McQueen et al., 1994; Norris et al., 1995). Our experiment and results suggested that cocktail party situations can be used as a new paradigm to study online lexical competition occurring during word identification. Interestingly enough, speech-in-speech comprehension situations offer a natural example where competition between linguistic information may occur and where its effects could be directly quantified through behavior.

To conclude, our experiment has clearly demonstrated that informational masking in multitalker speech-in-speech comprehension situations must be considered as a non-monolithic effect. In a babble sound, the availability of different types of linguistic information is modulated by its spectro-temporal saturation. In particular, our results showed that at lower talker-numbers ($N = 4$), both lexical and phonetic information were available from babble and competed with target identification, causing higher informational masking effects than in situations where the number of talkers was further increased. Hence, increasing the number of talkers in babble to over 4 resulted in a partial release from informational masking, rather than in a strengthening of it. Going back to psycholinguistic models of word recognition, this result supports the idea that using cocktail party situations offers an interesting experimental paradigm to study competition phenomena occurring during word identification although further experiments will have to be designed to specify more precisely the different factors that modulate the linguistic interference produced by babble as a background noise.

Acknowledgements

The authors would like to thank two anonymous reviewers for their very productive comments on a first version of the manuscript as well as R. Cusack for his help on a preliminary version of the manuscript and E. Veuillet for technical support at the onset of the study. The authors are very grateful for financial support of this project: the GDR CNRS 2213 for post-doctoral funding and the Rhone-Alpes region for an Emergence grant.

References

André-Obrecht, R., 1988. A new statistical approach for the automatic segmentation of continuous speech signals. IEEE Trans. on ASSP ~~36 (n1)~~.

Bacon, S.P., Viemeister, N.F., 1985. The temporal course of simultaneous tone-on-tone masking. J. Acoust. Soc. Amer. 78, 1231–1235.

Bolia, R.S., Nelson, W.T., Ericson, M.A., Simpson, B.D., 2000. A speech corpus for multitalkers communications research. J. Acoust. Soc. Amer. 107, 1065–1066.

Bregman, A.S., 1994. Auditory Scene Analysis. MIT, Cambridge.

Bronkhorst, A., 2000. The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. Acustica 86, 117–128.

Bronkhorst, A., Plomp, R., 1992. Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. J. Acoust. Soc. Amer. 92, 3132–3138.

Brungart, D.S., 2001a. Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Amer. 109, 1101–1109.

Brungart, D.S., 2001b. Evaluation of speech intelligibility with the coordinate response measure. J. Acoust. Soc. Amer. 109, 2276–2279.

Brungart, D.S., Simpson, B.D., Ericson, M.A., Scott, K.R., 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. J. Acoust. Soc. Am. 110, 2527–2538.

Brungart, D.S., Chang, P.S., Simpson, B.D., Wang, D., 2006. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. 120, 4007–4018.

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

Chi, T., Gao, Y., Guyton, M.C., Ru, P., Shamma, S., 1999. Spectro-temporal modulation transfer functions and speech intelligibility. J. Acoust. Soc. Am. 106, 2719–2732.

Danhauer, J.L., Leppler, J.G., 1979. Effects of four noise competitors on the California Consonant Test. J. Speech Hear. Disord. 44, 354–362.

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

Divenyi, P.L., 2004a. Speech Segregation by Humans and Machines. Kluwer Academic Publisher, Dordrecht, The Netherlands.

Divenyi, P.L., 2004b. The times of Ira Hirsh: multiple ranges of auditory temporal perception. Semin Hear. 25, 229–239.

Divenyi, P.L., Brandmeyer, A., 2003. The "cocktail-party effect" and prosodic rhythm: discrimination of the temporal structure of speech-like sequences in temporal interference. In: Solé, M.J., Recasens, D., Romero, J. (Eds.), Proc. 15th Internat. Congress of Phonetic Sciences, Barcelona, Spain, pp. 2777–2780.

Drullman, R., Bronkhorst, A., 2000. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. J. Acoust. Soc. Amer. 107, 2224–2235.

Duquesnoy, A.J., 1983. Effect of a single interfering noise or speech source on speech intelligibility. J. Acoust. Soc. Amer. 74, 739–743.

Egan, J.P., Carterette, E.C., Thwing, E.J., 1954. Some factors affecting multi-channel listening. J. Acoust. Soc. Amer. 26, 774–782.

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

Gaskell, M.G., Marslen-Wilson, W.D., 1997. Discriminating local and distributed models of competition in spoken word recognition. In: Shafto, M.G., Langley, P. (Eds.), Proc. Nineteenth Annual Conf. of the Cognitive Science Society, pp. 247–252.

Greenberg, S., 1995. The ears have it: the auditory basis of speech perception. In: Proc. Internat. Congress of Phonetic Sciences, Vol. 3, pp. 34–41.

Greenberg, S., Arai, T., 2001. The relation between speech intelligibility and the complex modulation spectrum. In: Proc. 7th Eurospeech Conf. on Speech Communication and Technology (Eurospeech-2001), pp. 473–476.

Greenberg, S., Hollenback, J., Ellis D., 1996. Insights into spoken language gleaned from phonetic transcription of the switchboard corpus. In: Proc. ICSLP'96, Philadelphia, USA.

Hawley, M.L., Litovsky, R.Y., Colburn, H.S., 1999. Intelligibility and localization of speech signals in a multi-source environment. J. Acoust. Soc. Amer. 105, 3436–3448.

Houtgast, T., Steeneken, J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Amer. 77, 1069–1077.

Karnebäck, S., 2001. Discrimination between speech and music based on a low frequency modulation feature. In: Proc. Eurospeech, pp. 1891–1894.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H&H hypothesis. In: Hardcastle, W., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publisher, Dordrecht, The Netherlands, pp. 403–409.

Luce, P.A., Pisoni, D.B., 1998. Recognizing spoken words: the neighborhood activation model. Ear Hearing 19, 1–36.

Luce, P., Pisoni, D., Goldinger, S., 1990. Similarity neighbourhoods of spoken words. In: Altmann, G. (Ed.), Cognitive Models of Speech Perception: Psycholinguistic and Computational Perspectives. MIT Press, Cambridge, USA, pp. 122–147.

Marslen-Wilson, W.D., 1987. Functional parallelism in spoken word-recognition. Cognition 25, 71–102.

Marslen-Wilson, W.D., 1990. Activation, competition, and frequency in lexical access. In: Altmann, G. (Ed.), Cognitive Models of Speech Perception: Psycholinguistic and Computational Perspectives. MIT Press, Cambridge, USA, pp. 148–172.

Marslen-Wilson, W.D., Moss, H.E., van Halen, S., 1996. Perceptual distance and competition in lexical access. J. Exp. Psychol.: Hum. Percept. Perform. 22, 1376–1392.

McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cogn. Psychol. 8, 1–86.

McQueen, J.M., Norris, D., Cutler, A., 1994. Competition in spoken word recognition: spotting words in other words. J. Exp. Psychol.: Learn. Mem. Cog. 20, 621–638.

Miller, G.A., 1947. The masking of speech. Psychol. Bull. 44, 105–129.

Monsell, S., Hirsh, K.W., 1998. Competitor priming in spoken word recognition. J. Exp. Psychol.: Learn. Mem. Cog. 24, 1495–1520.

Moss, H.E., McCormick, S., Tyler, L.K., 1997. The time-course of activation of semantic information during spoken word recognition: function precedes form. Lang. Cogn. Proc. 12, 695–733.

New, B., Pallier, C., Brysbaert, M., Ferrand, L., 2004. Lexique 2: a new French lexical database. Behav. Res. Meth. Instr. Comp. 36, 516–524.

Norris, D., 1994. Shortlist: a connectionist model of continuous speech recognition. Cognition 52, 189–234.

Norris, D., McQueen, J.M., Cutler, A., 1995. Competition and segmentation in spoken word recognition. J. Exp. Psychol.: Learn. Mem. Cog. 21, 1209–1228.

Peissig, J., Kollmeier, B., 1997. Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. J. Acoust. Soc. Am. 101, 1660–1670.

Pinquier, J., Rouas, J.L., André-Obrecht, R. 2003. A fusion study in speech/music classification. In: ICASSP'2003, Hong Kong.

Saberi, K., Perrott, D.R., 1999. Cognitive restoration of reversed speech. Nature 398, 760.

Scheirer, E., Slaney, M., 1997. Construction and evaluation of a robust multifeature speech/music discriminator. In: IEEE International Conference on Audio, Speech and Signal Processing, Munich, Germany, pp. 1331–1334.

Simpson, S.A., Cooke, M., 2005. Consonant identification in N-talker babble is a non-monotonic function of N (L). J. Acoust. Soc. Amer. 118, 2775–2778.

Steeneken, H.J.M., Houtgast, T., 1980. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Am. 67, 318–326.

Wood, N.L., Cowan, N., 1995. The cocktail party phenomenon revisited: attention and memory in the classic selective listening procedure of Cherry (1953). J. Exp. Psychol.: Learn. Mem. Cog. 21, 255–260.

Wrede, B., 2002. Modelling the effects of speech rate variation for automatic speech recognition. Doktor-Ingenieurin Dissertation, Universität Bielefeld, Technische Fakultät, 2002.

Zwitserlood, P., Schriefers, H., 1995. Effects of sensory information and processing time in spoken-word recognition. Lang. Cog. Proc. 10, 121–136.