The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception

Martin Cooke[a)]
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom

M. L. Garcia Lecumberri
Department of English Philology, University of the Basque Country, Paseo de la Universidad 5, 01006, Vitoria, Spain

Jon Barker
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom

(Received 8 March 2007; revised 8 October 2007; accepted 12 October 2007)

Studies comparing native and non-native listener performance on speech perception tasks can distinguish the roles of general auditory and language-independent processes from those involving prior knowledge of a given language. Previous experiments have demonstrated a performance disparity between native and non-native listeners on tasks involving sentence processing in noise. However, the effects of energetic and informational masking have not been explicitly distinguished. Here, English and Spanish listener groups identified keywords in English sentences in quiet and masked by either stationary noise or a competing utterance, conditions known to produce predominantly energetic and informational masking, respectively. In the stationary noise conditions, non-native talkers suffered more from increasing levels of noise for two of the three keywords scored. In the competing talker condition, the performance differential also increased with masker level. A computer model of energetic masking in the competing talker condition ruled out the possibility that the native advantage could be explained wholly by energetic masking. Both groups drew equal benefit from differences in mean F0 between target and masker, suggesting that processes which make use of this cue do not engage language-specific knowledge.
 [DOI: 10.1121/1.2804952]

PACS number(s): 43.71.Hw, 43.71.Es, 43.66.Dc [ARB] Pages: 414–427

I. INTRODUCTION

It is widely assumed that two kinds of processes play a part in decoding speech when other sound sources are present. First, general-purpose "signal-driven" processes are thought to help in creating an initial segregation of the auditory scene into components belonging to different acoustic sources (Bregman, 1990). Second, "knowledge-driven" processes which exploit prior knowledge of individual sources such as speech could be used to integrate the components into a coherent linguistic interpretation. The boundary between signal-driven and knowledge-driven processes is not clear, and there is some debate over the extent to which auditory scene analysis is capable of grouping the segregated components into coherent structures, or whether this is accomplished primarily by learned models of speech (Remez *et al.*, 1994).

Nearly all speech perception studies have focused on signal-driven processes. For example, in a competing talker experiment, listeners may be faced with sentence pairs whose mean fundamental frequency (F0) difference is the key variable. While these studies measure the effect of factors such as F0 differences on intelligibility, they say little about the role played by prior speech knowledge. One way to explore this latter factor is by comparing listener groups differing in the state of spoken language acquisition. Such nonhomogeneous listener groups are frequently chosen on the basis of native language (e.g., Florentine *et al.*, 1984) although it is also possible to vary prior linguistic experience by comparing perception in children and adults with the same native language (e.g., Hazan and Markham, 2004). If nonhomogeneous listener groups were to show similar intelligibility benefits of factors such as F0 differences, this would constitute strong evidence for the hypothesis that general auditory (or at least language-universal) processes are mainly responsible for F0-based source separation.

The presence of other sound sources results in masking of the "target" speech in ways summarized in Fig. 1. It is conventional to distinguish (i) energetic masking (EM), which occurs when components of the speech signal in some time-frequency region are rendered inaudible because of swamping by the masker, and (ii) informational masking (IM), which covers everything that reduces intelligibility

[a)]Author to whom correspondence should be addressed. Electronic mail: m.cooke@dcs.shef.ac.uk

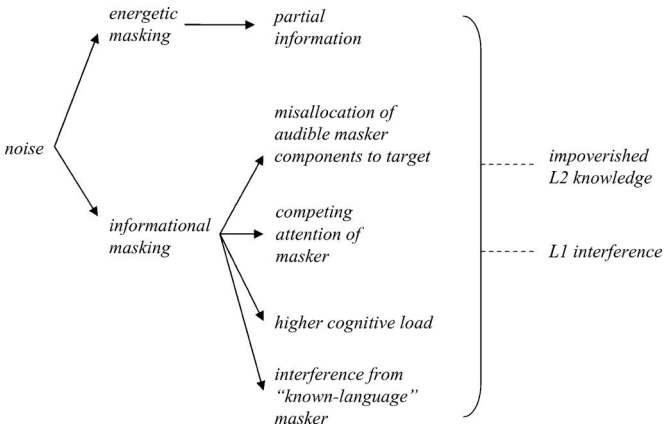

FIG. 1. Summary of potential masking effects for native and non-native listeners.

once energetic masking has been accounted for. Informational masking has multiple facets and is a catch-all term whose use reflects the current state of both conceptual and scientific uncertainty (Durlach, 2006).

The primary purpose of the current study was to investigate differential effects of energetic and informational masking on native and non-native speech perception and to relate the findings to the relative roles of signal- versus knowledge-driven processes in the perception of speech.

Energetic masking leads to a loss of signal components so that partial information has to be used to interpret the speech signal. Fortunately, speech is sufficiently redundant that, for many tasks, only a relatively small proportion of the time-frequency plane has to be "glimpsed" to support high levels of intelligibility (Cooke, 2006; Barker and Cooke, 2007).

Several studies have compared native and non-native or bilingual speech perception in stationary noise, which can be considered to be a pure energetic masker (Florentine *et al.*, 1984; Hazan and Simpson, 2000; Bradlow and Bent, 2002; van Wijngaarden *et al.*, 2002; Garcia Lecumberri and Cooke, 2006; Rogers *et al.*, 2006). While the languages used varied and the tasks ranged from consonant identification in short tokens to keyword identification in sentences, all these studies demonstrated that natives outperform non-natives in noisy conditions. However, for those studies which compared performance in quiet (or low noise) and high noise conditions, estimates of the relative size of the native advantage differed. While one study (Bradlow and Bent, 2002) found that the native advantage remained constant as noise level increased, another (Garcia Lecumberri and Cooke, 2006) demonstrated an increased native advantage in the high noise condition.

We now consider the various potential elements of informational masking listed in Fig. 1. One, misallocation, denotes situations where the listener uses audible elements from the masker to arrive at an incorrect identification of the target, or, equivalently, assigns target elements to the masker, again resulting in an error. Informational masking of speech has most often been studied using maskers which are themselves composed of speech material (Carhart *et al.*, 1969; Brungart, 2001; Freyman *et al.*, 2004). When a single competing talker is present, whole words from the masker can be reported as belonging to the target (Brungart, 2001). However, misallocation could, in principle, apply to units of any size. In particular, patches of acoustic information (e.g., bursts, formant transitions, or frication) might be wrongly attributed. For this reason, any maskers containing speech could produce informational masking through misallocation. In fact, significant IM effects have been demonstrated for *N*-talker babble over a wide range of values for *N* (Simpson and Cooke, 2005). Since misallocation can apply to units smaller than words or phonemes, it might also result in the reporting of a sound or word which is not present in either the target or masking speech. For example, the aspiration following a plosive could be interpreted as the voiceless glottal fricative "h."

A second component of IM comes from the higher cognitive load which results when processing a signal containing multiple components. If both target and masker might contain important information, it is reasonable to suppose that processing resources are allocated to both. A related facet of IM is the failure to attend to the target in the presence of a competing speech masker. The role of differences in fundamental frequency, vocal tract size, and spatial cues in determining which of two competing sentences is attended to has been studied (Darwin and Hukin, 2000). If attention is based on limited resources (e.g., Kahneman, 1973), then a higher cognitive load may well result in difficulties in tracking the target source.

A further aspect of informational masking is dependent on whether the language of the masking speaker is known to listeners. A number of recent studies have demonstrated that the language of the masker can affect the intelligibility of the target sentence (Rhebergen *et al.*, 2005; Garcia Lecumberri and Cooke, 2006; Van Engen and Bradlow, 2007). Rhebergen *et al.* found worse speech reception thresholds for Dutch sentences presented in competing Dutch speech than when the competitor material was Swedish. Using a consonant in vowel context identification task, Garcia Lecumberri and Cooke (2006) showed that monolingual English listeners performed better when the language of a competing speaker was Spanish rather than English, whereas Spanish listeners with English as their second language (L2) were equally affected by maskers in both languages. Van Engen and Bradlow (2007) demonstrated that for native English listeners, English sentence intelligibility was better when the noise consisted of two-talker Mandarin Chinese babble than when it was composed of two-talker English babble. These results suggest that a masking talker using a language known to the listener increases informational masking relative to one using an unknown language, perhaps due to the engagement of language-specific decoding processes for both masker and target, which in turns increases cognitive load.

In contrast to energetic masking, the role of informational masking in non-native speech processing has received little attention to date. A number of researchers have compared native and non-native or bilingual performance using maskers composed of speech babble (Mayo *et al.*, 1997; Cutler *et al.*, 2004; Garcia Lecumberri and Cooke, 2006), which

is known to be capable of inducing IM (Simpson and Cooke, 2005). While all three studies mentioned IM-related factors such as native language (L1) interference, none of them supported a quantitative assessment of the size of any native advantage in IM, and the tasks involved were not designed with IM in mind.

Consequently, a key goal of the current study was to compare native and non-native performance using a competing talker task which is known to induce extensive informational masking (Brungart, 2001). The competing talker situation arises frequently in speech communication, so it is of interest to determine whether informational masking effects cause significantly more problems for non-native listeners. Brungart presented listeners with pairs of sentences added at a range of target-to-masker ratios (TMRs). Sentences were drawn from the CRM corpus (Bolia *et al.*, 2000), which consists of sentences with a simple structure such as "ready baron go to green 4 now" or "ready charlie go to red 3 now." The call sign ("baron," "charlie") acts as a keyword to identify which of the sentences of the pair the listener is to attend to, and the task usually involves reporting the color and the digit of the attended sentence. Brungart varied the availability of potential cues for separating the sentence pairs and estimated the amount of informational masking in each case. In one condition, listeners were asked to identify keywords from a target talker when the same talker was used as a masker. This condition provides few cues to separate the two utterances and produced large amounts of informational masking, which was especially evident when target and masker sentences were mixed at the same rms level. Listeners were able to use a level difference cue even in the same talker condition to improve keyword identification scores. In other conditions, the target and masking talkers were of the same gender or of differing genders, providing cues such as voice quality and F0 differences which contributed to a reduction in informational masking.

The current study employed a task similar to that used by Brungart to explore the role of informational masking in non-native speech perception. Speech material was drawn from the "Grid" corpus (Cooke *et al.*, 2006), which consists of 1000 sentences from each of 34 talkers. Sentences have a particularly simple six-word form such as "place white at B 4 now" and "lay green with N 8 again." The availability of many utterances and individual talkers in the corpus also allowed the effect of factors such speech rate, F0 differences, and individual speaker intelligibility to be compared across the native and non-native groups.

Experiment 1 measured native and non-native listeners' keyword identification scores for Grid sentences in quiet and in three levels of speech-shaped noise (SSN). The primary goal was to derive an estimate of pure energetic masking to enable the contribution of EM in experiment 2 to be estimated in order to better assess the role of IM. A further goal was to determine whether an increasing native advantage with greater amounts of energetic masking found in our earlier study with VCV tokens (Garcia Lecumberri and Cooke, 2006) carries over to the sentence material of the Grid corpus. Sentences contain a wider range of phonetic realizations, including intersegmental effects, admit more variation in speaking rate and prosodic structure, and invoke a higher cognitive load than isolated VCVs. By using simple sentences, our aim was to introduce some natural variation while restricting the use of high-level knowledge. Although it is known that native listeners are better able to exploit higher-level knowledge such as syntactic context contained in sentence-level material in processing noisy speech (Mayo *et al.*, 1997; van Wijngaarden *et al.*, 2004), the Grid corpus was felt to minimize demands on higher-level processing since all utterances are syntactically, semantically, and pragmatically equal and sufficiently short to reduce memory loading. In addition, the total lexicon used in the corpus is a collection of 51 very common words, of which only 39 act as keywords, minimizing non-native listener disadvantage due to deficits in the L2 lexicon. Our goal in using Grid sentences was similar to that of Meador *et al.* (2000), who used semantically unpredictable sentences composed of words likely to be known to non-native listeners. A further aim was to compare the intelligibility of multiple talkers for the two listener groups. Grid contains 34 talkers of both genders with a variety of accents, and differences in the intelligibility ranking of talkers might be expected based on the native listeners' richer knowledge of talker differences.

Experiment 2 compared informational masking in natives and non-natives and measured keyword identification scores for target sentences in the presence of a competing sentence in conditions where the availability of cues to the separability of the sentence pair was systematically varied in the manner of Brungart (2001).

II. EXPERIMENT 1: SENTENCES IN STATIONARY NOISE

A. Methods

1. Participants

The non-native group consisted of 49 native speakers of (European) Spanish. All were students at the University of the Basque Country studying English as a foreign language (age range: 20–25, mean: 21.2 years). They were enrolled in a one-semester course in English Phonetics in the second year of a four-year BA degree in English Language and Literature. All students had attained the level of the Cambridge Advanced Examination. Students received course credit for participating in the listening tests. The results of 7 of the 49 Spanish listeners were excluded from the analysis of experiment 1 based on their performance in experiment 2 (see Sec. III A 1).

Results for native listeners were derived from an earlier study on speaker identification in noise (Barker and Cooke, 2007), which employed similar stimuli. Twenty English listeners took part in the Barker and Cooke study, 18 of which also participated in experiment 2 of the current study. For this reason, results from the common subset of 18 were extracted for comparison with non-natives in experiment 1.

2. Speech and noise materials

Speech material was drawn from the Grid corpus (Cooke *et al.*, 2006), which consists of six word sentences such as "lay red at H 3 now." In experiment 1, colors, letters, and digits acted as keywords. Four choices of color ("red," "white," "green," and "blue"), 25 letters of the English alphabet (excluding "W" due to multisyllabicity), and ten spoken digits ("one" to "nine" and "zero") were available. All 34 talkers (18 male, 16 female) who contributed to the Grid corpus were used.

Listeners heard utterances without noise and in three stationary noise conditions created by the addition of SSN whose long-term spectrum was the average of sentences in the Grid corpus. Noise was added at token-wise signal-to-noise ratios (SNRs) of 6, 0, and −6 dB. Spanish listeners identified 60 sentences in each of the four conditions while native listeners had been tested in an earlier study across a larger range of SNRs and heard 100 sentences in each condition (Barker and Cooke, 2007).

Two considerations led to the choice of SNRs for the non-native group. First, since one goal of the experiment was to provide a means of estimating the effect of energetic masking in experiment 2, the noise levels had to be chosen such that they would result in a similar degree of energetic masking as found in the competing talker conditions in that experiment. Since a competing talker provides around 6–8 dB less masking than does stationary noise when presented at the same SNR (Miller, 1947; Festen and Plomp, 1990), it was necessary to avoid the use of extremely low SNRs in experiment 1. A second reason for choosing these SNRs was on the basis of estimates of the expected non-native performance disadvantage, derived by extrapolating from our earlier study (Garcia Lecumberri and Cooke, 2006). There, native listener performance in stationary noise was at the same level as that of non-native listeners in a competing talker condition, suggesting a non-native deficit of around 6–8 dB. While Garcia Lecumberri and Cooke used VCVs rather than sentences, the lack of strong contextual cues in the current task suggests that a similar non-native deficit might result for both types of stimuli. By applying this estimate to the full SNR-intelligibility relation for natives in the current task provided by Barker and Cooke (2007), the SNR values of 6, 0, and −6 dB were predicted to provide a representative mapping of the non-native SNR-intelligibility relation.

To enable sufficient representation of speech material from different talkers, individual listeners heard different sets of sentences in each condition. For the natives, each listener heard a different set of sentence/talker combinations drawn at random from the Grid corpus. For the non-native group, ten different sets of sentences/talker combinations were extracted from the corpus at random, and each listener was randomly assigned to one of the ten sets.

3. Procedure

The non-native group was tested at the University of the Basque Country. Stimulus presentation and response collection was under computer control. Listeners were asked to identify the color, letter, and digit spoken and entered their results using a conventional computer keyboard in which four of the nonletter/digit keys were marked with colored stickers. Listeners were familiarized with the stimuli and the task by identifying an independent practice set of 60 sentences in quiet prior to the main set. Stimuli were blocked according to noise level, and the order of the blocks was randomized across listeners.

Native listeners had been tested individually in an IAC single-walled acoustically isolated booth using Sennheiser HD250 headphones at the University of Sheffield. At the University of the Basque Country, participants were tested in groups of 15–20 in a quiet laboratory using Plantronics Audio-90 headphones. The difference in the two stimulus presentation setups in the two countries was shown to be nonsignificant in a previous study involving the perception of VCV tokens in noise (Garcia Lecumberri and Cooke, 2006).

B. Results

Listener responses were scored separately for each of the three keywords. Due to near-ceiling performance in quiet and the low noise condition for the native group, scores were converted to rationalized arcsin units (RAU; Studebaker, 1985) for both statistical analyses and graphical displays. Figure 2 shows RAU-transformed keyword scores for the two groups together with the native advantage (N-NN), expressed as a difference in RAUs. As expected, native listeners performed better in all conditions. For the color and number keywords, the native advantage increased with background noise level (from 6 to 14 RAUs for the color keyword and from 12 to 27 RAUs for the number keyword). On average, the native advantage was least for the color keyword and greatest for the letter keyword. Separate repeated measures ANOVAs with one within-subjects factor (noise level) and one between-subjects factor (nativeness) were performed for each of the three keywords, confirming the clear effects of noise level and nativeness in each case. The interaction noise level × nativeness was significant for color $[F(3,56)=2.85,\ p<0.05,\ \eta^2=0.13]$ and number $[F(3,56)=12.6,\ p<0.001,\ \eta^2=0.40]$ but not for letter $(p=0.29)$.

These results suggest that non-native listeners suffer more from the effects of increasing stationary background noise when identifying certain keywords in simple sentences. Interestingly, the native advantage for the most difficult keyword (letter) did not increase with noise level.[1] It is not clear why non-native listeners did not suffer more for the highly confusable set of spoken letters, though it is notable that the native advantage in quiet was already large.

As mentioned earlier, the native listeners of the Barker and Cooke (2007) study were exposed to a larger range of SNRs and a greater number of tokens in each noise condition than the non-native group. To determine whether the greater exposure to the task was beneficial for the natives, the full range of conditions from the earlier study was analyzed for both within-condition learning effects and across-condition

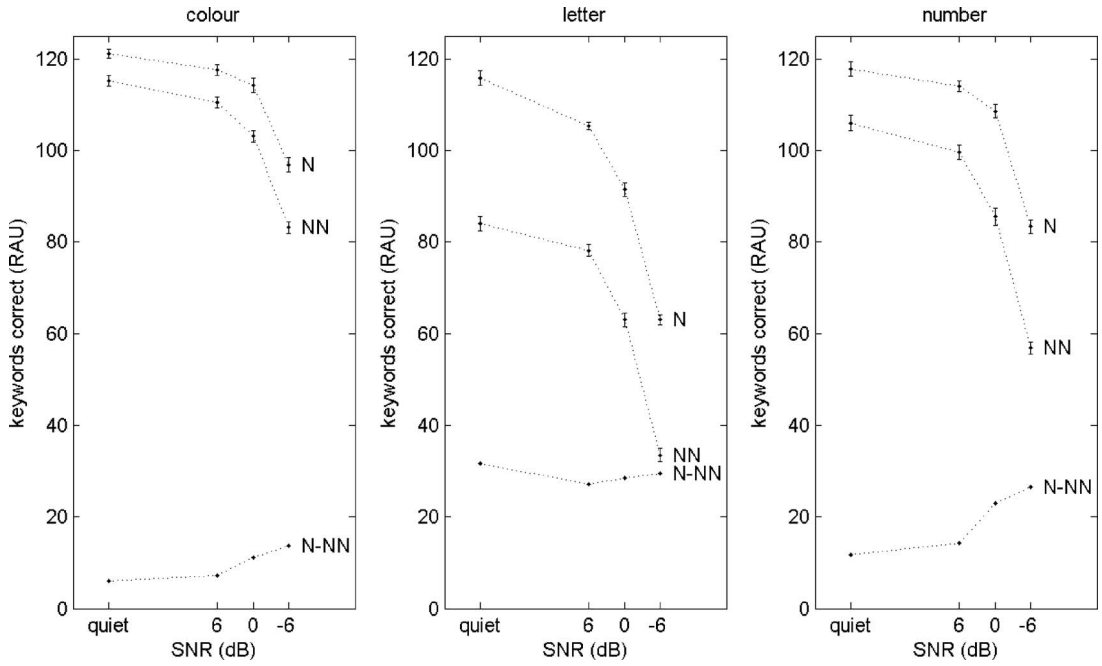

FIG. 2. Native (N) and non-native (NN) keyword identification scores in rationalized arcsine units (RAUs) for quiet and in three levels of speech-shaped noise for color, letter, and number keywords. The native advantage (N-NN in RAUs) is also shown. Error bars here and elsewhere denote ±1 standard errors.

order effects. No such learning effects were found,[2] suggesting that the differences reported here were due to differences in the two groups of listeners.

C. Acoustic and speaker analyses

To further explore the origins of energetic masking suffered by listeners, a number of additional analyses involving the factors of gender, speech rate, and talker were performed. For these analyses, results are presented as percentage keywords correct, transformed to RAUs.

1. Gender

The top row of Fig. 3 shows the breakdown of intelligibility by speaker gender for native and non-native listeners in the four conditions of experiment 1. Both listener groups showed a similar pattern in each condition. Utterances from neither gender proved more intelligible than the other in quiet and low levels of noise. However, female speakers were more intelligible by approximately equal amounts for both groups in higher levels of noise. A two-way ANOVA (gender × nativeness) at each SNR confirmed a significant intelligibility advantage for female talkers at both 0 dB[$F(1,58)=6.5$, $p<0.05$, $\eta^2=0.10$] and −6 dB[$F(1,58)=57.1$, $p<0.001$, $\eta^2=0.50$]. It also confirmed the lack of a gender by nativeness interaction in any of the four conditions, suggesting that both groups benefitted equally from the more intelligible female talkers.

2. Speech rate

The lower row of Fig. 3 illustrates the effect of speech rate on intelligibility. Since utterances have the same number of words (and nearly the same number of syllables), speech rate is approximately in inverse proportion to utterance duration. Utterances were split into three equal-sized groups based on duration, and keyword identification scores for the fastest and slowest groups were compared. Overall, listeners scored significantly better for the slower utterances ($p<0.001$) in quiet and at all SNRs. In the quiet condition, a small but significant interaction between duration and nativeness was present [$F(1,58)=5.0$, $p<0.05$, $\eta^2=0.08$]. Here, the native group showed no benefit of increased duration, probably because their performance was near ceiling.

3. Individual talkers

Figure 4 shows scatterplots of mean talker intelligibility scores for the native versus non-native groups in quiet and each of the three levels of speech-shaped noise. Correlations between native and non-native scores are also shown. Apart from quiet, all correlations are statistically significant ($p<0.01$ for the 6 dB condition, $p<0.001$ for the 0 and −6 dB conditions).

In the quiet condition, most of the native scores are at or near "ceiling" levels (no listener errors are made for 21 of the 34 talkers). However, the non-native group found certain talkers more difficult than others. For example, talkers m28 and f33 are outliers (defined here as lying more than 2 s.d. from the mean). These talkers are the only two with a Scottish accent in the Grid corpus, and it is perhaps not surprising that nonstandard accents are problematic for non-native listeners, who lack the exposure to a range of regional variation. Indeed, dialectal/idiolectal variation can contribute to poor non-native perception (Strange, 1995) and the intelligibility of unfamiliar accents is correlated with the nativeness of the listener (Ikeno and Hansen, 2006).

As conditions become increasingly adverse, native and non-native judgments of talker difficulty converge, as demonstrated by the increase in correlation from 0.44 at 6 dB SNR to 0.80 at −6 dB SNR. For example, in the latter con-

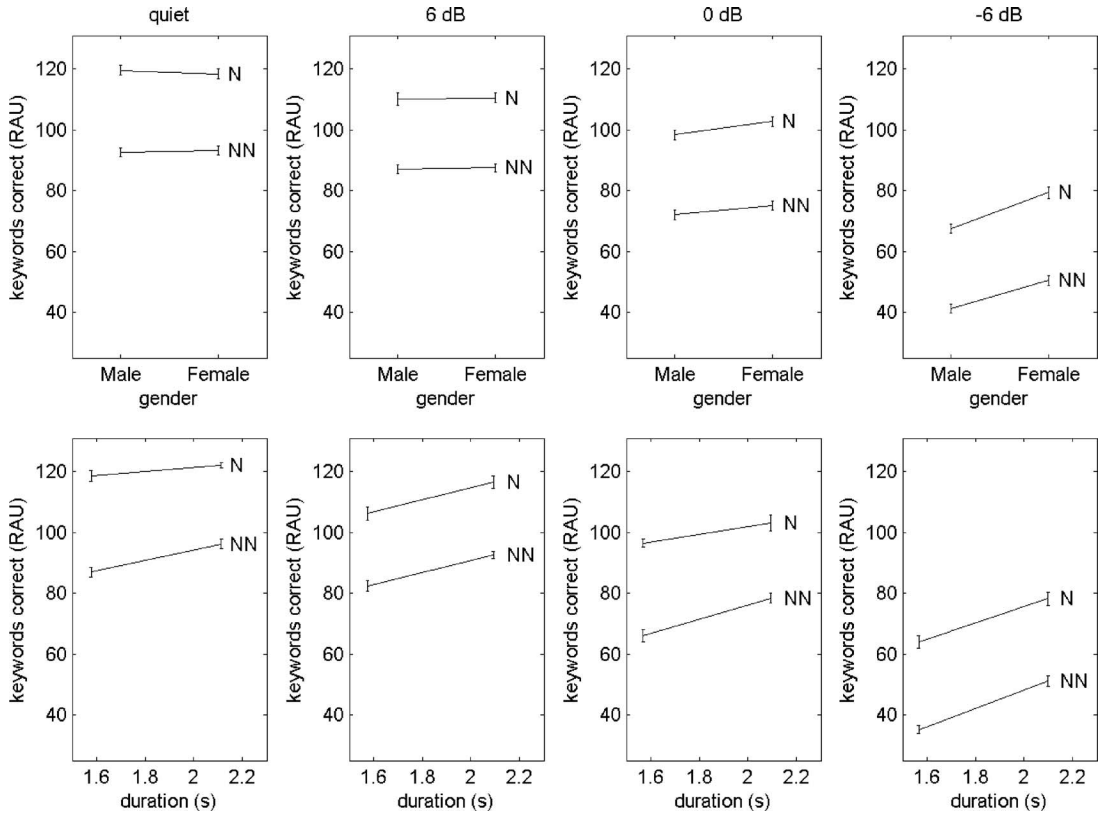

FIG. 3. Effect of speaker gender (top) and speech rate (bottom) on intelligibility for native and non-native listeners in the four conditions of experiment 1. Scores are averaged over color, letter, and number keywords and transformed to RAUs. For speech rate, the ordinate represents the mean duration of the fastest and slowest thirds of the utterance set.

dition, while talkers such as f7 and fl8 are highly intelligible for natives and non-natives, talker m1 is problematic for both groups. The increasing similarity for the two listener groups in noise suggests that accent and other idiosyncratic speaker-related information is rendered less salient by energetic masking, while talker characteristics which promote robust speech cues that resist masking are useful to native and non-native listeners, although not necessarily equally so.

D. Summary and discussion

Experiment 1 measured the effect of pure energetic masking on the two listener groups and revealed that the non-native group suffered more from increasing levels of noise for two of the three keywords in the simple sentences used here. An analysis of gender effects in experiment 1 demonstrated a similar pattern of increasing intelligibility of female talkers in the high noise conditions for both listener groups. The factors that underpin the higher mean female intelligibility in noise for this corpus are not known but appear to be equally useful to both native and non-native listeners, suggesting that it is relatively low-level acoustic differences such as higher formant frequencies or language-independent differences in speaking style rather than language-specific factors which govern the female intelligibility advantage. Bradlow *et al.* (1996) and Hazan and Markham (2004) also found that females were more intelligible for speech presented in quiet conditions. In the current study, the masker had a long-term spectrum derived from both male and female talkers, so it is possible that the higher center of gravity of female speech (caused both by a higher mean F0 and higher formant frequencies) led to some release from masking relative to male talkers.

Non-native listeners benefitted from a slower speech rate in quiet and all noise levels, while natives benefitted in all but the quiet condition. A slower speaking rate can help in two ways. First, in absolute terms, it leads to a greater "visibility" of the target speech. If an informative acoustic feature is masked at one instant, it may be "glimpsed" at a later instance with a probability inversely proportional to the speaking rate. Of course, slower speaking rates do not lengthen all sounds equally, so the increased glimpsing opportunities may not reduce misidentifications across all phonemes. Second, a slower speaking rate results in a slower information rate and thus reduced attentional load for higher-level tasks such as lexical retrieval. While the increased visibility of the target will help both natives and non-natives in noise, it seems plausible that non-natives will benefit from anything which increases the available processing time due to the greater complexity of speech perception in a second language (Gass, 1997).

One striking finding of experiment 1 was that while native and non-natives found different individual talkers more or less intelligible in quiet and low noise conditions, both groups tended to agree on an intelligibility ordering of talkers in noisier conditions. A similar finding was reported for automatic speech recognition scores in noise for this corpus (Barker and Cooke, 2007). This suggests that in situations of low noise, where detailed acoustic information pertaining to

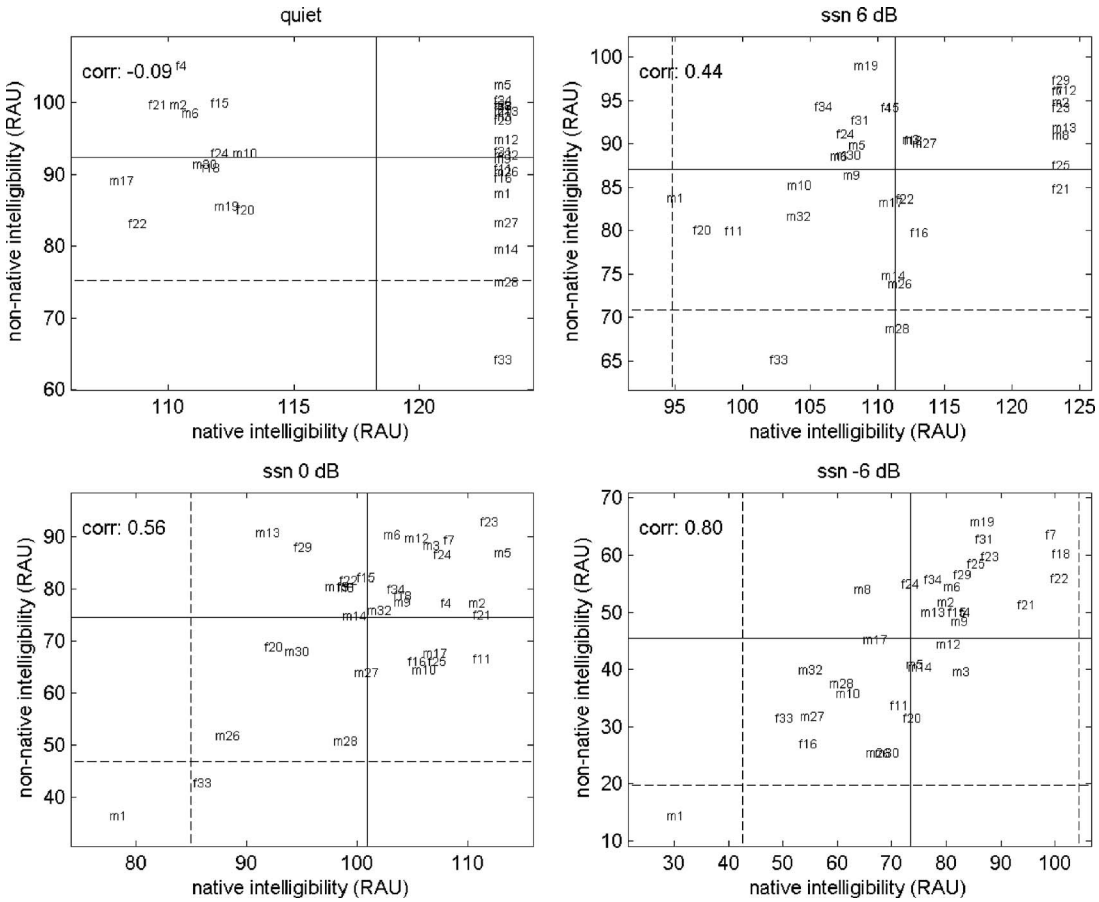

FIG. 4. Intelligibility of individual talkers for native and non-native listeners in the four conditions of experiment 1. Solid lines indicate mean intelligibilities across all talkers for the two listener groups, while dotted lines locate ±2 s.d. from the mean (these are plotted where the value is within the range of talker scores). Male and female talkers are identified by m and f, respectively, and the numbers distinguish different talkers in the Grid corpus. Scores computed as for Fig. 3.

individual talkers is available, previous experience of speech variation produced by factors such as differing accents are dominant in producing the native advantage. Native listeners are able to draw upon a richer knowledge base in interpreting the signal. However, in the presence of high levels of noise, these knowledge-driven factors appear to give way to more general acoustic factors that make individual talkers more resistant to noise, since both groups find the same talkers difficult or easy to recognize. This might be seen as a generalization of the result for different genders. A talker with a "peaky" spectrum (i.e., with energy concentrated at the information-bearing formant frequencies) will resist energetic masking more than one whose spectral profile is more diffuse. It appears that both native and non-native listeners benefit from the more informative distribution of glimpses of the target which result.

III. EXPERIMENT 2: COMPETING TALKERS

A. Methods

1. Participants

The same listeners (47 non-native, 18 native) who took part in experiment 1 participated in experiment 2. However, 7 of the Spanish participants were deemed to be responding randomly in the most difficult conditions (mean keyword identification of 2% in the most difficult condition compared with a mean of 47% for the rest of the non-native group). Consequently, their results were excluded from the analysis of both experiments.

2. Speech materials

As in experiment 1, utterances were drawn from the Grid corpus (Cooke *et al.*, 2006). Sentences were paired to be approximately equal in duration and added at six different TMRs: 6, 3, 0, −3, −6, −9 dB, chosen on the basis of Brungart (2001). The target sentence always contained the keyword "white." The letter and digit keywords always differed in the target and masker. Following Brungart (2001), sentence pairs were split into three subconditions: "same talker," "same gender," and "different gender." There were 20 sentence pairs in each of the three subconditions, leading to a total of 60 pairs at each of the 6 TMRs.

3. Procedure

Stimulus presentation was as described for experiment 1. Listeners reported the letter and digit spoken by the target, identified as the speaker producing the keyword "white." Listeners were familiarized with the task through an independent practice set of 60 utterance pairs. In the main part of the experiment, each TMR constituted a block. Presentation order of the six blocks was randomly chosen for each lis-

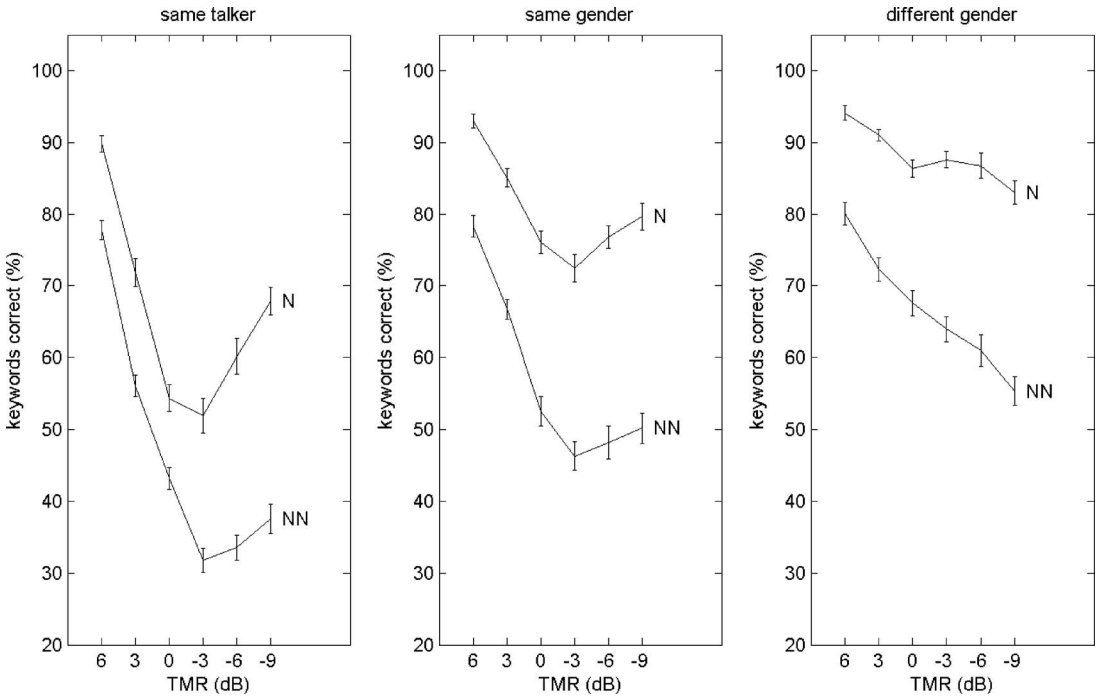

FIG. 5. Native and non-native keyword identification scores in the two-talker conditions.

tener. Within each block, the 60 utterance pairs were presented in a random order. Consequently, the same talker, same gender, and different gender utterance pairings were mixed within blocks. To prevent listeners from using absolute level as a cue to the target utterance, presentation level was randomly roved over a 9.5 dB range from stimulus to stimulus.

B. Results

Results are presented as percentage keywords correct scored for the (letter, digit) keyword pair. As for experiment 1, percentages were converted to RAUs for statistical analysis. However, this produced essentially identical outcomes as for raw percentages, so results are displayed in terms of percentages for ease of interpretation.

Figure 5 presents native and non-native keyword identification performance as a function of TMR in the three simultaneous talker subconditions. The pattern of results for native listeners was similar to that found by Brungart (2001). Listeners had least difficulty in identifying target keywords when the masking talker was of a different gender, and had most difficulty when the same talker was used for target and masker. The strongly nonmonotonic pattern as a function of TMR in the same talker and same gender conditions was also found by Brungart, and demonstrates the beneficial effect of level differences between target and masker in helping to assign keywords to the target speaker. In the different gender case, other cues are sufficiently strong to render level differences unnecessary. These results confirm the usefulness of the Grid corpus in studies of informational masking in speech. Overall, scores for the non-native group followed the same pattern.

For the current study, whose focus is on differences in native and non-native performance in a task designed to produce large amounts of informational masking, the main feature of interest in the results is the large native advantage at all TMRs and in all subconditions. The native advantage ranges from 12 to 15 percentage points at the most favorable TMR and reaches 30 percentage points at the least favorable TMR.

A repeated measures ANOVA with two within-subjects factors (TMR and sentence pairing condition) and one between-subjects factor (nativeness) showed that the three-way interaction of TMR × nativeness × sentence pairing was not significant ($p>0.5$). However, all two-way interactions were significant [TMR × nativeness: $F(5,54)=8.7$, $p<0.001$, $\eta^2=0.45$; sentence pairing × nativeness: $F(2,57)=5.4$, $p<0.01$, $\eta^2=0.16$; sentence pairing × TMR: $F(10,49)=30.2$, $p<0.001$, $\eta^2=0.86$]. The first of these (TMR × nativeness) confirms that non-native listeners are more seriously disadvantaged at adverse TMRs than natives. The second interaction (sentence pairing × nativeness) suggest that the pattern of native advantage is different for each sentence pair type (same talker, same gender, different gender), although the effect is small. The remaining interaction (sentence pairing × TMR) arises from the differing non-monotonic behavior with TMR across the three conditions. The analysis also confirmed highly significant effects of TMR [$F(5,54)=127$, $p<0.001$, $\eta^2=0.92$], speaker pairing condition [$F(2,57)=369$, $p<0.001$, $\eta^2=0.93$], and nativeness [$F(1,58)=131$, $p<0.001$, $\eta^2=0.69$].

C. Acoustic analyses

As for experiment 1, analyses were performed to determine how the two listener groups responded to low-level factors. Here, fundamental frequency differences between the target and masker sentences and absolute duration were examined.

1. Fundamental frequency differences

A difference in fundamental frequency (F0) between two simultaneous sentences has long been known to improve intelligibility (Brokx and Nooteboom, 1982; Bird and Darwin, 1998). F0 difference is considered a primitive source separation cue which is independent of the type of source and hence ought to be equally beneficial to listeners, regardless of their first language.

For each sentence pair employed in experiment 2, the mean instantaneous difference in F0 (measured in semitones) was computed for all frames where both utterances were voiced. F0 information and binary voicing decisions were computed at 10 ms intervals automatically using an autocorrelation approach implemented in PRAAT (Boersma and Weenink, 2005). Using the same approach as was applied for duration in experiment 1 (Sec. II C 2), sentence pairs were split into three equal-sized groups based on F0 difference and the lower and upper groups compared. The three subconditions (same talker, same gender, and different gender) were analyzed separately to prevent the differing extent of mean F0 differences in the three subconditions from masking any differences in the lower and higher terciles.

The effect of F0 differences on keyword identification in each of the two-talker conditions is shown in the upper panel of Fig. 6. Significant effects were found in all three subconditions [same talker: $F(1,58)=7.79$, $p<0.01$, $\eta^2=0.12$; same gender: $F(1,58)=40.4$, $p<0.001$, $\eta^2=0.41$; different gender: $F(1,58)=6.51$, $p<0.05$, $\eta^2=0.10$]. None of the interactions with nativeness were significant, suggesting that F0 differences were equally beneficial for the two groups. The smallest effect of F0 differences was found in the different gender condition. All utterance pairs in the different gender condition had large F0 differences, ranging from a mean of around 5 semitones in the lower tercile to nearly an octave in the higher tercile. These findings suggest that, for both listener groups, the effects of F0 differences reach a ceiling at around 5 semitones, a result in line with the findings (for native listeners) of Darwin *et al.* (2003), who employed similar sentences and near-identical competing talker conditions.

2. Speech rate

To determine the effect of speech rate on keyword identification in the two-talker conditions, an analysis similar to that described in Sec. II C 2 was performed. Since utterances were paired by similar duration, this analysis is not of speech rate differences between the utterances in a pair, but rather examines the effect of absolute speech rate. The results shown in the lower panel of Fig. 6 combine durational information across the three two-talker subconditions since very similar patterns were observed in analyses of each subcondition. The interaction between duration and nativeness was significant [$F(1,58)=7.2$, $p<0.01$, $\eta^2=0.11$]. Non-natives benefitted significantly from a slower overall speaking rate [$F(1,58)=46.5$, $p<0.001$, $\eta^2=0.45$] while natives did not ($p>0.2$).

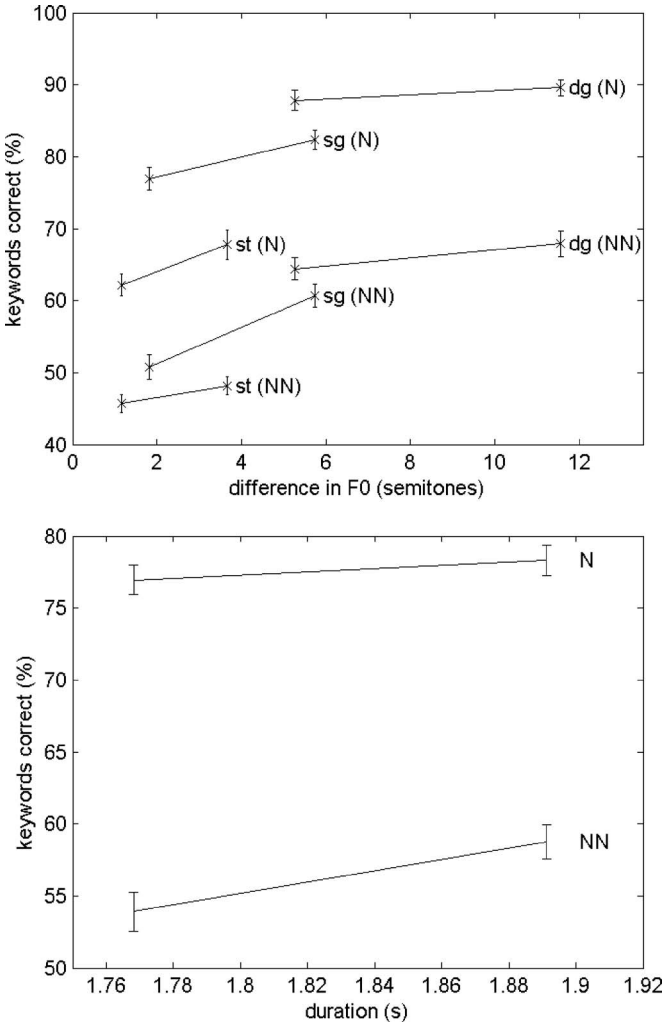

FIG. 6. Upper panel: Effect of fundamental frequency differences in the two-talker conditions. Keyword identification scores for natives and non-natives in the three subconditions (sg=same gender, st=same talker, dg=different gender) are presented for subsets of utterance pairs in the lower and upper tercile of F0 differences. Lower panel: Effect of absolute duration on keyword identification in the two-talker conditions.

D. Quantifying the degree of informational and energetic masking in the two-talker situation

At issue in the current study is the origin of the native advantage in the two-talker case. A competing talker provides relatively little energetic masking at the TMRs used here (Brungart *et al.*, 2006) and informal listening suggests that the letter-digit pair from both target and masker are usually clearly audible in the two-talker signals. Following Brungart (2001), one way to assess the extent of informational masking in the two-talker task is to examine the proportion of listener responses which were present in the masker rather than the target utterance. These errors might be considered to result from informational masking. Figure 7 partitions listener responses into three categories: keywords correctly reported, i.e., present in the target (black), errors where the keywords reported were from the masking talker (midgray) and keywords not contained in either utterance of the pair (light gray). In general, native listeners make fewer "masker confusions" than non-native listeners in the three speaker-pairing subconditions. At the higher TMRs, the na-

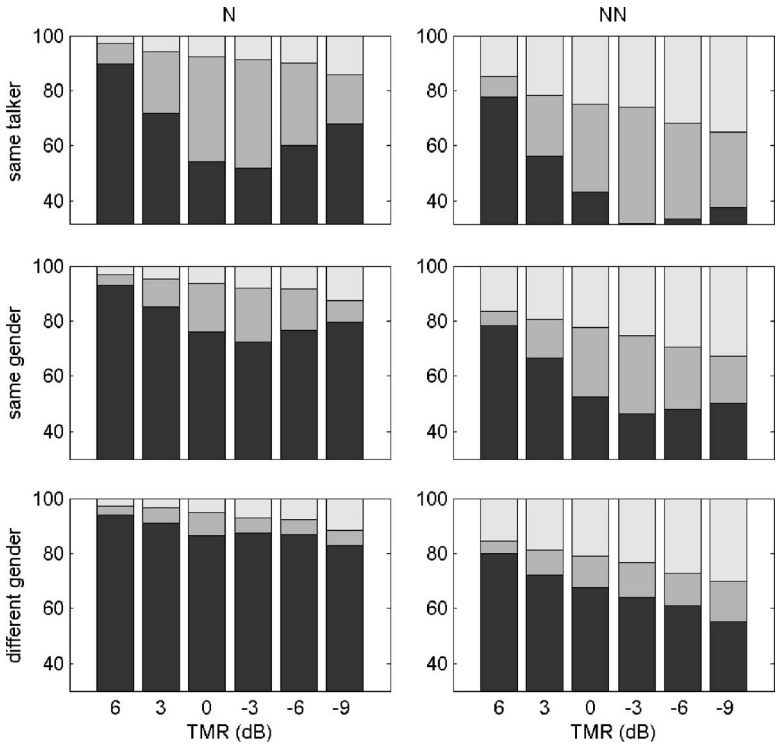

FIG. 7. Proportions of keywords from the target utterance (black) and from the masker (mid gray). The residual (light gray) shows the proportion of responses which were not part of the target or masker.

tive advantage is slight but increases to around 9 percentage points at the lowest TMR. The three subconditions show different degrees of native advantage. The same talker condition shows the least difference between the two groups while the greatest difference occurs in the same gender subcondition. This may reflect the relative ease of the different gender task for both listener groups, reducing the scope for any native advantage.

The difference in the proportion of keywords reported that were present in neither the target nor masker increases monotonically with decreasing TMR, from approximately 12 to 20 percentage points, with a similar increase in each of the three speaker-pairing subconditions. It is tempting to ascribe this type of error solely to energetic masking, but, while it is likely that many of these errors do originate in EM, it is also possible that listeners sometimes combine acoustic cues from the target and masker to "invent" a third sound. This is a form of informational masking by misallocation. It is more likely to occur in the Grid corpus, which contains the highly confusable spoken alphabetic letters, than in corpora such as CRM (Bolia *et al.*, 2000; Brungart, 2001), which used a restricted range of color keywords in the equivalent position. In support of this notion, listeners make 5.2 times as many errors in reporting spoken letters not present as in reporting digits, twice as many as would be expected on the basis of the relative number of response alternatives. Consequently, the proportion of keywords present in neither target nor masker cannot be seen as a reliable measure of energetic masking in the competing talker situation.

An alternative approach to quantifying the extent of energetic masking in experiment 2 is to extrapolate from the results of the pure energetic masking conditions of experiment 1. The technique is based on a glimpsing model (Cooke, 2006), which assumes that intelligibility in a pure energetic masking situation is a function of both prior speech knowledge and the availability of glimpses of the target speech in regions not dominated by the masker. For each listener group, prior speech knowledge is the same in both experiments. Consequently, it is possible to use the pure energetic masking conditions of experiment 1 to estimate the relationship between intelligibility and the proportion of the spectrotemporal plane glimpsed, and to use this relationship to quantify energetic masking effects in experiment 2 by measuring the glimpse proportion at each TMR.

One potential problem with the use of the glimpse proportion metric is that it ignores the distributional characteristics of glimpses in the time-frequency plane. One would expect different glimpse distributions to result from the stationary and competing talker maskers of experiments 1 and 2, although to some extent the distribution of foreground speech glimpses will be similar regardless of the masker type due to the relatively sparse concentration of energy in harmonics and formants. To determine whether glimpse proportion is a good predictor of intelligibility for the different glimpse distributions resulting from the maskers of experiments 1 and 2, a "missing-data" automatic speech recognition system modified to handle glimpses (Cooke *et al.*, 1994; 2001) was used to identify keywords in the stimuli of both experiments. The automatic speech recognizer (ASR) scores at the four SNRs of experiment 1 (measured as raw percentages) and the six TMRs of experiment 2 are shown in the upper panel of Fig. 8. Since only four noise levels were used in experiment 1, piecewise linear interpolation was used to estimate the glimpse-intelligibility relation across the entire range of glimpse proportions. Figure 8 shows that the ASR scores in the two experiments are very similar at the same glimpse proportions, being within 1 percentage point at those glimpse proportions where they overlap, apart from the low-

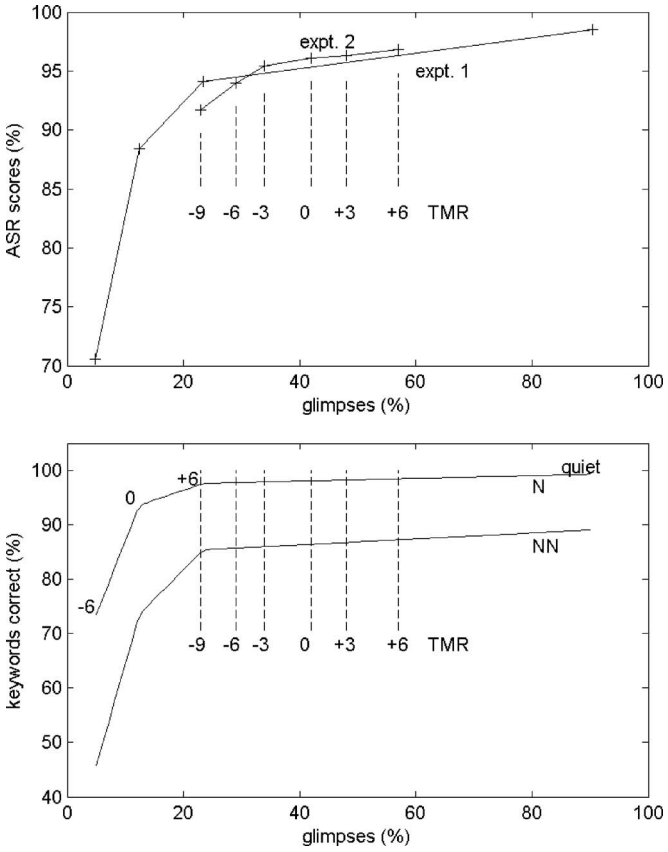

FIG. 8. Upper panel: Automatic speech recognition scores based on glimpse recognition for the stimuli of experiments 1 and 2. Rather than being expressed in terms of SNR (experiment 1) or TMR (experiment 2), recognition scores are plotted as a function of the mean glimpse percentage on which recognition was based. Lower panel: Solid lines (after pairwise linear interpolation between the four SNRs indicated) depict glimpse proportion vs intelligibility (scored as the mean identification rate of the letter and digit keywords) derived from experiment 1, for native and non-native listeners. Vertical lines indicate the measured glimpse percentages for the six TMR conditions of experiment 2.

est TMR value of −9 dB where the scores differ by 2.5%. Consequently, it appears that glimpse proportion alone can be used to mediate between intelligibility reductions due to energetic masking in the two experiments.

The lower panel of Fig. 8 plots intelligibility versus glimpse percentage for the native and non-native groups based on the results of experiment 1. To permit comparison of the two experiments, experiment 1 was rescored based on the two keywords (letter and number) identified in experiment 2. Also plotted in Fig. 8 are the glimpse percentages at the six TMRs used in experiment 2. The intelligibilities at the points where the vertical lines intersect the two curves are the estimates of the energetic masking effect for natives and non-natives in the competing talker experiment. These estimates suggest that the energetic masking effect of a competing talker is rather small, even at the lowest TMR used, and varies over a narrow range across TMRs. Indeed, the predicted keyword identification scores, if energetic masking were the only factor operating, range from 97.5 to 98.5 for natives and from 85.0 to 87.3 for non-natives.

A further estimate of the effect of energetic masking is provided by the ASR scores shown in the upper panel of Fig. 8. A comparison of the ASR scores with native listeners' identification rates in experiment 1 (Fig. 8, lower panel) suggests that the glimpsing model applied in the ASR system is about 3 percentage points worse than listeners on average. Consequently, the energetic masking effect suggested by ASR scores almost certainly overestimates the masking effect on listeners.

A promising alternative approach to isolating the energetic masking component was recently developed by Brungart *et al.* (2006). Like the approach described here, their technique uses a model of energetic masking to determine which portions of the target are audible, but instead of using glimpse counts or ASR scores based on glimpses, a resynthesis technique was used to reconstruct those parts of the signal which resist energetic masking, and the resulting signal is scored by listeners. They also found that energetic masking plays a relatively small role in the overall masking effect in the two-talker situation.

E. Summary and discussion

The native listener results of experiment 2 confirm the findings of Brungart (2001) and extend them to a more extensive and challenging corpus. More important, experiment 2 provides what we believe to be the first test of the "non-native cocktail party," where listeners had to identify keywords spoken by the target talker in the presence of a competing talker uttering very similar material. The speech-on-speech masking condition is known to provoke large amounts of informational masking. Here, native listeners scored between 10 and 30 percentage points better than non-natives. While it is difficult to determine precisely how much of this deficit was due to energetic masking, several analyses suggested that informational masking played by far the dominant role. Even after accounting for the effect of energetic masking on the two groups, there remains a native advantage of up to 10 percentage points due to reduced informational masking. In fact, analyses based on a computational model of energetic masking estimate a higher deficit of perhaps 20 percentage points. The results of experiment 2 suggest that non-native listeners are more adversely affected than natives by informational masking in multiple talker situations, and that the native advantage increases as the relative level of the masking talker increases.

Further acoustic analyses of the two-talker experiment demonstrated that both groups drew equivalent benefits from differences in the mean fundamental frequencies of the two simultaneous sentences. This is a very strong indication that the processes which lead to intelligibility improvements with increases in F0 difference precede the engagement of native-language-specific speech processes, since if the latter were involved in exploiting F0 differences (for example, by taking advantage of the more "visible" target speech harmonics which might results from F0 differences), one would expect to see a greater benefit for native talkers in conditions of large F0 difference.

Speech rate also played a part in the two-talker conditions. Non-natives identified substantially more keywords in slower utterances than in the more rapid utterances, while the

effect of speech rate was marginal for native listeners. As in the pure energetic masking case, this probably reflects the advantages of a slower information rate in a task which makes great attentional and cognitive demands.

The analysis of keyword confusions (Fig. 7), where listeners reported tokens from the masking source, are of particular interest. Non-natives found the same gender condition particularly confusing but, intriguingly, reported similar numbers of confusions as natives in the two positive TMRs of the same talker condition, and indeed reported fewer confusions than the natives at a TMR of 0 dB in that condition. Results in the different gender condition were intermediate between the other two conditions. To make sense of these findings, it is helpful to consider the cues which listeners might use to separate speakers in the three conditions.

First, in the same talker condition, cues such as differences in level and F0 as well as continuity of formants and harmonics are available. It seems that when the target is least masked, native and non-native listeners misallocate masker components to the target at about equal rates, suggesting that both groups are equally able to exploit level and F0 differences. In the same gender condition, additional cues are available. These fall into two classes: those that are language-universal (e.g., differences in voice quality, vocal tract length) and those which are language-specific (e.g., differences in accent and other speaker idiosyncrasies). Since the native group is best placed to take advantage of the latter type of cue [e.g., Ikeno and Hansen (2006) demonstrated that native listeners are better at detecting and classifying accents], it is not surprising that this group makes fewer background confusions. In the different gender condition, speaker differences are more extreme. Consequently, the native advantage seen in the same gender condition is somewhat reduced. It is unlikely that native listeners benefitted from the fact that multiple talkers were presented in a mixed order. Bradlow and Pisoni (1999) showed that native and non-native listeners drew similar advantages from having a consistent talker.

The differences discussed above apply to the situations where the target speaker is dominant. However, non-natives are much more likely to report keywords from the masker when the masker is dominant. In principle, there are at least two strategies that listeners could use to solve the two-talker problem. One would involve the use of cues such as F0 differences to "track" the separate talkers through time from the color keyword to the appropriate letter-digit combination. An alternative approach is to extract speaker "tags" from the color keyword and match these against the appropriate letter-digit keywords. Indeed, informal listening to two-talker utterances suggests that a more sophisticated form of tagging is possible, whereby listeners use "tags" from the color keyword belonging to the masker in order to eliminate letter-digit keywords produced by the masker. Such a strategy is the obvious one to use at lower TMRs when the masker is dominant.

If tracking were the dominant method, one might expect similar scores for the two listener groups, since both show similar benefits of F0 differences and tracking would seem to be a linguistically universal process. Since the difference in confusion scores in the negative TMR conditions is so great, it is more likely that a tagging strategy is dominant in this task. Given that in solving the two-talker problem listeners have to not only pick out the weaker target keywords but also to detect speaker cues (such as gender, mean F0, voice quality, accent) based on the color keyword in order to decide which letter-digit keywords to report, it is clear that in a tagging approach the non-native group has a double disadvantage because of their less-rich models of language variation.

Finally, caution is required in interpreting the target/masker confusions as wholly the result of informational masking. Since energetic masking plays a dual role in the two-talker case (color identification followed by letter-digit identification), it is difficult to ascribe all of the target/masker confusions to informational masking. It may well be that both letters and digits are audible and that some combination is reported so that the results appear to favor informational masking, but if the decision on which combination to report is based on a partially audible color keyword, then energetic masking is partly responsible for the results.

IV. GENERAL DISCUSSION

The two experiments reported in this paper demonstrate that in a task involving the identification of keywords in simple sentences spoken by native English talkers, Spanish listeners are more adversely affected than English listeners by increases in masker level for both stationary noise and competing speech maskers.

In principle, non-native listeners suffer both because of impoverished knowledge of the second language, and due to interference from their first language (Trubetzkoy, 1939; Strange, 1995). It is of interest to consider how these factors interact with effects of masking as listed in Fig. 1 to determine possible origins of the non-native deficit in noise.

In the case of energetic masking, native listeners presumably perform well due to their extensive experience of speech and in particular the effects of masking on the signal. Since there are multiple redundant cues to important phonetic distinctions such as voicing (Lisker, 1986), native listeners may have learned which cues survive in different noise conditions. On the other hand, non-native listeners have far less exposure to the second language and indeed may have virtually no experience of hearing the L2 in noisy conditions, so one might expect to see a differential effect of energetic masking, with non-native listeners suffering more in adverse conditions.

Of the multiple causes of informational masking, non-native listeners might be expected to suffer more than natives from target/masker misallocation. The accuracy of "sorting" audible components into speech hypotheses is likely to be higher for listeners with richer knowledge of the target language, which can be used to prevent false rejections and acceptances. For example, a listener whose knowledge of English is restricted to one specific accent may be less able to assign speech sounds from other accents to the target or

the background source (McAllister, 1997; Strange, 1995). Similarly, influences from the non-native L1 may create further difficulties in allocation of sound components, particularly at the phonemic level required to report the spoken letters in experiment 2. Cues in the target which do not conform to L1 categories may be wrongly allocated to the masker.

The other facets of IM listed in Fig. 1 may also be responsible for reduced performance amongst non-native listeners. The two-talker task creates a high cognitive load even for native listeners, and there is evidence that some aspects of processing a foreign language are slower than in processing a native language (Callan *et al.*, 2004; Mueller, 2005; Clahsen and Felser, 2006). Speech segregation processes requiring tracking and focus of attention may also affect non-natives adversely. For example, if listeners lack knowledge about English stress-timed rhythm, they are missing what may be a useful cue in segregating and tracking competing speech sources. Likewise, interference from L1 expectations of intonational contours might affect tracking.

V. CONCLUSIONS

The two experiments reported in this paper attempted to quantify the effect of energetic and informational masking on native and non-native listeners. English and Spanish listener groups identified keywords in simple sentences presented in stationary speech-shaped noise and in the presence of a competing talker speaking a similar sentence. Both conditions induced significantly more errors in the non-native group. A computer model suggested that the effect of energetic masking on the two groups could not account for the large native advantage in the competing talker conditions. It can be concluded that non-native listeners suffer a large performance deficit due to informational masking relative to native listeners.

Just as comparisons involving speech and nonspeech (e.g., music) sources can be used to distinguish the roles of general auditory from speech-specific processes, studies comparing listener populations with different native languages can be used to distinguish those parts of the speech interpretation process which make use of language-specific prior knowledge from those which are speech specific but language independent. In the current study, both groups derived equal benefit from differences in mean fundamental frequency between the target and masking talker, suggesting that segregation of speech using fundamental frequency cues has no language-specific component. Further studies will determine which other potential cues for understanding speech in noise act independently of prior linguistic knowledge.

ACKNOWLEDGMENTS

This work was supported by grants from the Spanish Ministry of Science and Technology, the Basque Government (9/UPV 00103.130-13578/2001) and the University of Sheffield Research Fund. We thank Jonny Laidler and Balakrishnan Kolluru for useful comments on the manuscript as well as three anonymous reviewers and Ann Bradlow for their insightful comments.

[1]In raw percentage terms, the native advantage increased for all three keywords (colors: 1% in quiet to 10% at −6 dB; letters: 16% to 31%; numbers: 4% to 25%).

[2]Within-condition learning effects: The mean difference between second and first half token presentations across conditions and listeners was −0.01 percentage points. Across-condition: The correlation between listeners' standardized (z) scores and condition order was insignificant (correlation =0.009; $p=0.89$).

Barker, J., and Cooke, M. P. (**2007**). "Modelling speaker intelligibility in noise," Speech Commun. **49**, 402–417.

Bird, J., and Darwin, C. J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr, London), pp. 263–269.

Boersma, P., and Weenink, D. (**2005**). "Praat: Doing phonetic by computer," version 4.3.14 (computer program), http://www.praat.org. Last accessed 10 August 2007.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bradlow, A. R., and Bent, T. (**2002**). "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**, 272–284.

Bradlow, A. R., and Pisoni, D. B. (**1999**). "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors," J. Acoust. Soc. Am. **106**, 2074–2085.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech. I. Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**, 255–272.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organisation of Sound* (MIT Press, Cambridge, MA).

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

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

Callan, D. E., Jones, J. A., Callan, A. M., and Akahane-Yamada, R. (**2004**). "Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models," Neuroimage **22**, 1182–1194.

Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**, 694–703.

Clahsen, H., and Felser, S. (**2006**). "How native-like is non-native language processing?," Trends Cogn. Sci. **10**, 564–570.

Cooke, M. P. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

Cooke, M. P., Barker, J., Cunningham, S. P., and Shao, X. (**2006**). "An audio-visual corpus for speech perception and automatic speech recognition," J. Acoust. Soc. Am. **120**, 2421–2424.

Cooke, M. P., Green, P. D., and Crawford, M. D. (**1994**). "Handling missing data in speech recognition," *Proceedings of the Third International Conference on Spoken Language Processing*, Yokohama, Japan, pp. 1555–1558.

Cooke, M. P., Green, P. D., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and uncertain acoustic data," Speech Commun. **34**, 267–285.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (**2004**). "Patterns of English phoneme confusions by native and non-native listeners," J. Acoust. Soc. Am. **116**, 3668–3678.

Darwin, C. J., and Hukin, R. W. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–977.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

Durlach, N. (**2006**). "Auditory masking: Need for improved conceptual structure," J. Acoust. Soc. Am. **120**, 1787–1790.

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

Florentine, M., Buus, S., Scharf, B., and Canevet, G. (**1984**). "Speech reception thresholds in noise for native and non-native listeners," J. Acoust. Soc. Am. **75**, s84 (abstract).
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.
Garcia Lecumberri, M. L., and Cooke, M. P. (**2006**). "Effect of masker type on native and non-native consonant perception in noise," J. Acoust. Soc. Am. **119**, 2445–2454.
Gass, S. M. (**1997**). *Input, Interaction and the Second Language Learner* (Lawrence Erlbaum Associates, Mahwah, NJ).
Hazan, V., and Markham, D. (**2004**). "Acoustic-phonetic correlates of talker intelligibility in adults and children," J. Acoust. Soc. Am. **116**, 3108–3118.
Hazan, V., and Simpson, A. (**2000**). "The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects," Lang Speech **43**, 273–294.
Ikeno, A., and Hansen, H. L. (**2006**). "Perceptual recognition cues in native English accent variation: Listener accent, perceived accent, and comprehension," *International Conference on Acoustics Speech and Signal Processing*, Toulouse, France, pp. 401–404.
Kahneman, D. (**1973**). *Attention and Effort* (Prentice-Hall, Englewood Cliffs, NJ).
Lisker, L. (**1986**). "Voicing in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees," Lang Speech **29**, 3–11.
Mayo, L. H., Florentine, M., and Buus, S. (**1997**). "Age of second-language acquisition and perception of speech in noise," J. Speech Lang. Hear. Res. **40**, 686–693.
McAllister, R. (**1997**). "Perceptual foreign accent: L2 users' comprehension ability," in *Second Language Speech: Structure and Process*, edited by A. James and J. Leather (Mouton de Gruyter, New York).
Meador, D., Flege, J. E., and MacKay, I. R. (**2000**). "Factors affecting the recognition of words in a second language," Bilingualism: Lang. Cognit. **3**, 55–67.
Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.
Mueller, J. L. (**2005**). "Electrophysiological correlates of second language processing," Second Lang. Res. **21**, 152–174.
Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (**1994**). "On the perceptual organization of speech," Psychol. Rev. **101**, 129–156.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118**, 1274–1277.
Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., and Abrams, H. B. (**2006**). "Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing," Appl. Psycholinguist. **27**, 465–485.
Simpson, S., and Cooke, M. P. (**2005**). "Consonant identification in N-talker babble is a non-monotonic function of N," J. Acoust. Soc. Am. **118**, 2775–2778.
Strange, W. (**1995**). *Speech Perception and Linguistic Experience* (York, Timonium, MD).
Studebaker, G. A. (**1985**). "A "rationalized" arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Trubetzkoy, N. (**1939**). *Principles of Phonology (Grundzüge der Phonologie)* (University of California Press, Berkeley).
Van Engen, K. J., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**, 519–526.
van Wijngaarden, S. J., Bronkhorst, A. W., Houtgast, T., and Steeneken, H. J. M. (**2004**). "Using the Speech Transmission Index for predicting non-native speech intelligibility," J. Acoust. Soc. Am. **115**, 1281–1291.
van Wijngaarden, S. J., Steeneken, H. J. M., and Houtgast, T. (**2002**). "Quantifying the intelligibility of speech in noise for non-native listeners," J. Acoust. Soc. Am. **111**, 1906–1916.