A glimpsing model of speech perception in noise

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

(Received 18 March 2005; revised 19 December 2005; accepted 19 December 2005)

Do listeners process noisy speech by taking advantage of "glimpses"—spectrotemporal regions in which the target signal is least affected by the background? This study used an automatic speech recognition system, adapted for use with partially specified inputs, to identify consonants in noise. Twelve masking conditions were chosen to create a range of glimpse sizes. Several different glimpsing models were employed, differing in the local signal-to-noise ratio (SNR) used for detection, the minimum glimpse size, and the use of information in the masked regions. Recognition results were compared with behavioral data. A quantitative analysis demonstrated that the proportion of the time–frequency plane glimpsed is a good predictor of intelligibility. Recognition scores in each noise condition confirmed that sufficient information exists in glimpses to support consonant identification. Close fits to listeners' performance were obtained at two local SNR thresholds: one at around 8 dB and another in the range −5 to −2 dB. A transmitted information analysis revealed that cues to voicing are degraded more in the model than in human auditory processing. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2166600]

PACS number(s): 43.66.Ba, 43.71.Es, 43.72.Dv, 43.66.Dc [DOS] Pages: 1562–1573

I. INTRODUCTION

Visually, objects are frequently identified based on partial views due to occlusion by other objects. However, the occlusion metaphor is not so obvious in hearing since the contributions from different acoustic objects combine additively in the sound mixture reaching the ears. Consequently, many perceptual and engineering studies have examined the separation of a target signal from a collection of sources that make up the background. Computational attempts at speech separation have been inspired by auditory scene analysis (Bregman, 1990; Cooke and Ellis, 2001), source independence (Comon, 1994; Hyvarinen *et al.*, 2001) and prior source models (Varga and Moore, 1990; Gales and Young, 1993). In practice, none of these approaches has been successful in extracting complex signals such as speech in everyday adverse conditions representative of those faced by listeners, since masking at the auditory periphery complicates the estimation of the energy contribution of the target speech source in each time–frequency region.

Two characteristics of speech signals motivate a different approach to understanding how speech might be recognized in noise. First, since speech is a highly modulated signal in time and frequency, regions of high energy are typically *sparsely distributed*. Consequently, the spectrotemporal distribution of energy in a mixture contains regions that are dominated by the target speech source, even at quite adverse signal-to-noise ratios (SNRs). In such regions, the "noisy" energy observations are very close to those in clean speech, rendering the problem of energy separation unnecessary. Second, the information conveyed by the spectrotemporal energy distribution of clean speech is *redundant*, as demonstrated by numerous studies of the intelligibility of speech after spectral filtering (Fletcher, 1953; Warren *et al.*, 1995; Lippmann, 1996; Kasturi *et al.*, 2002), temporal gating or modulation (Miller and Licklider, 1950; Strange *et al.*, 1983; Gustafsson and Arlinger, 1994), or spectrotemporal impoverishment (Drullman, 1995; Shannon *et al.*, 1995). Redundancy allows speech to be identified based on relatively sparse evidence.

Sparseness and redundancy give rise to an account of speech perception in noise based on the use of "glimpses" of speech in spectrotemporal regions where it is least affected by the background. Figure 1 depicts the regions of a modeled spectrotemporal excitation pattern (STEP; Moore, 2003) dominated by speech in the presence of three different maskers at a global SNR (i.e., SNR measured over the entire token) of −6 dB. Details of the STEP computation are given in Sec. IV B 1. The speech token, shown in the upper panel, is the syllable /ɑrɑ/ spoken by a male. The three panels in the middle row show, from left to right, the token masked by a single talker, eight-talker babble, and speech-shaped noise, respectively. The lower panels depict potential glimpses of the speech target in each of these masking conditions. In this figure, glimpses are defined as those spectrotemporal regions, where the speech energy exceeds that of the masker by at least 3 dB, although the effect of other values of this threshold are reported later in the paper. A substantial proportion of regions are unaffected by the masker, even at an adverse global SNR. Figure 1 also suggests why global SNR is not, on its own, a good predictor of intelligibility. Even though the global SNR is identical in all masking conditions, the typical glimpse size differs. Many studies have demonstrated that intelligibility at a fixed global SNR depends on the type of masker used. A single competing talker or amplitude-modulated noise is a far less effective masker than

[a)]Telephone: +44 114 2221822; Fax: +44 114 2221810. Electronic mail: m.cooke@dcs.shef.ac.uk

 0001-4966/2006/119(3)/1562/12/$22.50

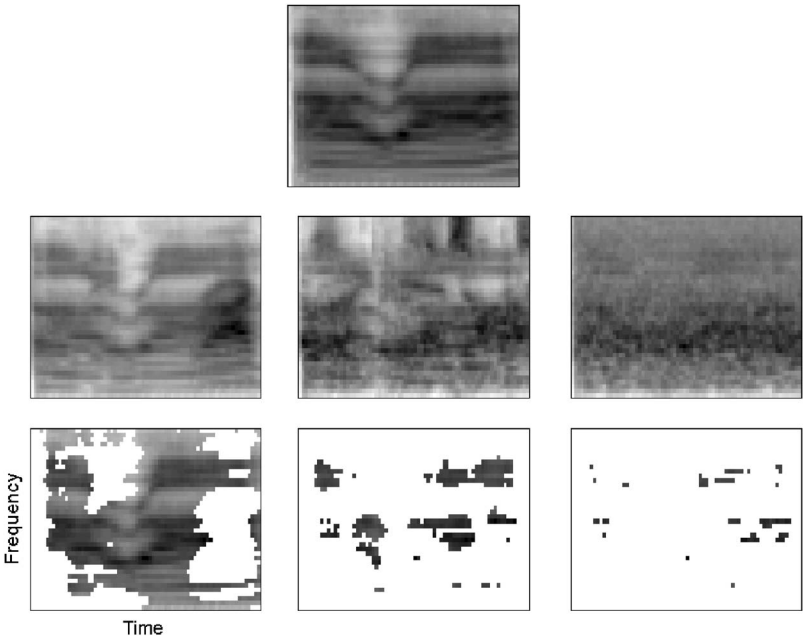

FIG. 1. An illustration of potential glimpses of a short speech token in three masking conditions. Top: spectrotemporal excitation pattern (STEP) for the speech token /ɑrɑ/. Middle: STEP representations for the mixture of the clean speech token with speech from a single competing talker (left), eight talkers (middle), and speech-shaped noise (right). Bottom: regions of the STEP, where the energy of the speech token exceeds that of the masker by at least 3 dB.

multispeaker babble or speech-shaped noise (Miller, 1947; Festen and Plomp, 1990; Scott *et al.*, 2004; Simpson and Cooke, 2005).

In the single competing talker condition, much of the evidence for /ɑrɑ/ survives in large chunks since the masker has significant spectral and temporal energy modulations. Fewer elements remain for the eight-talker masker due to filling in of the energy dips. However, the structures that survive in this example retain evidence of formant movement in the /ɑ/ to /r/ transition as well as intermittent evidence of vowel formants. In contrast, glimpses in the speech-shaped noise background are sparse and small at this SNR due to the absence of significant spectrotemporal modulations in the masker.

The notion that listeners use glimpses to identify speech in noise has been invoked by several researchers (Miller and Licklider, 1950; Howard-Jones and Rosen, 1993; Assmann and Summerfield, 1994; Culling and Darwin, 1994; Assmann, 1996) and has received renewed attention in recent years (Buss *et al.*, 2003; Buss *et al.*, 2004; Cooke, 2003; Assmann and Summerfield, 2004; Alcantara *et al.*, 2004; Freyman *et al.*, 2004).

Miller and Licklider (1950) measured the intelligibility of words in sentences whose waveforms had been interrupted by gating them on and off at a range of modulation frequencies. Across a broad range of interruption rates centered around 15 Hz, word identification scores were almost at the level of uninterrupted speech, in spite of the loss of 50% of the waveform. Miller and Licklider suggested that listeners were able to identify sentences by piecing together glimpses of the target speech available in the uninterrupted portions. In the Miller and Licklider study, glimpses were synchronous in that the complete spectrum was available. Using a "checkerboard" noise masker, Howard-Jones and Rosen (1993) investigated whether listeners could identify consonants in conditions that required the integration of asynchronous glimpses. By varying the spectral extent of each "checkerboard square," Howard-Jones and Rosen demonstrated that listeners were able to integrate glimpses occurring asynchronously, but only for wide spectral regions, accounting for around a quarter to a half of the frequency region of importance for speech perception. Buss *et al.* (2003) studied the effect on spondee identification of amplitude modulation (AM) coherence of either a noise masker or the speech signal filtered into nine narrow bands. They found that the identification of masked AM speech did not depend on the coherence of the modulating waveforms across frequency bands, suggesting that listeners are capable of piecing together relatively narrow spectral glimpses of speech occurring asynchronously. Buss *et al.* (2004) confirmed and extended this finding using a consonant identification task in 16 frequency bands.

The studies of Miller and Licklider (1950), Howard-Jones and Rosen (1993), and Buss *et al.* (2003, 2004) provided a temporal window of glimpsing opportunities whose duration was of the order of a phoneme. Since the most energetic regions of speech occur primarily during voiced episodes, it is reasonable to assume that temporal modulations at the mean voiced–unvoiced alternation rate of speech lead to phoneme-sized intervals of dominance, at least in those spectral regions occupied by formants. Other investigators have proposed that glimpsing opportunities of rather shorter durations may be exploited by listeners. Culling and Darwin (1994) suggested that waveform interactions that give rise to envelope modulation or "beating" provide brief glimpsing opportunities that allow listeners to identify one or other member of a simultaneous vowel pair. They showed that a vowel classifier operated on a sliding 10 ms temporal window at the output of a filterbank analysis could account for listeners' identification rates for vowels whose fundamental frequencies (F0s) were similar enough to produce significant beating. Assmann (1996) extended the Culling and Darwin model to vowels embedded in CVC syllables. Both studies concluded that brief glimpses of the entire spectral profile that occur as the result of waveform interactions can support the identification of vowels with close F0s. However,

the example in Fig. 1 suggests that naturally-produced speech in a range of background sources does not result in many temporal regions where information across the entire frequency spectrum is glimpsed. Instead, brief glimpses of *partial* spectral information do occur.

The intelligibility of speech resynthesized from partial information was measured by Roman *et al.* (2003) in order to assess the performance of an algorithm that used location cues to separate the contributions from two or three sources positioned in a number of spatial configurations. They found that resynthesis from partial spectrotemporal information led to large speech intelligibility improvements over the unprocessed mixture. Similarly, a recent study by Brungart *et al.* (submitted) measured the intelligibility of speech resynthesised from fragments similar to those depicted in Fig. 1. They demonstrated that intelligibility remained at high levels, even for putative glimpses of a speech target in a background composed of four competing talkers. The Roman *et al.* (2003) and Brungart *et al.* (submitted) studies suggest that sufficient information may exist in glimpses to support human speech perception. However, the task of identifying speech synthesized from partial, but clean, information is somewhat different from that faced by listeners, who have the additional problem of identifying which parts of a noisy signal should be treated as glimpses of the target speech.

The current study adopted a complementary approach to that of Brungart *et al.* by comparing the outputs of a computational model of glimpsing with listeners' performance on the same task (Simpson and Cooke, 2005). This approach allowed different models for the detection and integration of glimpses to be evaluated with respect to behavioral data. The glimpsing model employed missing-data algorithms derived from those used in robust automatic speech recognition (Cooke *et al.*, 1994; Cooke *et al.*, 2001), reviewed in Sec. II. Section III describes the speech in noise task and behavioral results, and provides a quantitative analysis of glimpsing opportunities afforded by a number of different maskers. The glimpse detection and identification model is described in Sec. IV.

II. REVIEW OF AUTOMATIC SPEECH RECOGNITION WITH MISSING DATA

Missing-data automatic speech recognition (ASR) was introduced by Cooke *et al.* (1994) as a technique for handling partial data in robust ASR. It has subsequently been applied as a computational model of vowel identification (de Cheveigné and Kawahara, 1999), in the recognition of sine-wave speech (Barker and Cooke, 1997) and as a model of narrow band speech perception (Cunningham, 2003). It has been used as a component of engineering systems for robust ASR to handle additive noise (Drygajlo and El-Maliki, 1998; Raj *et al.*, 1998) and reverberation (Palomaki *et al.*, 2002). Missing-data recognition also forms the core of a probabilistic decoder for multiple acoustic sources (Barker *et al.*, 2005).

Although missing-data techniques can be used in many ASR architectures, in this section it is shown how they can be applied to the most commonly used approach, namely hidden Markov models (HMMs). A HMM is typically used to model units of speech such as phones, triphones, syllables or words, and consists of a number of states, each of which models some segment of the acoustic waveform. States are linked by directed arcs that indicate allowable state transitions, and each transition has a probability associated with it. The relationship between HMM states and the speech waveform is not fixed in advance. Instead, the model learns both the state transition probabilities and probability distributions for each state from a large amount of training data. Each state probability distribution represents an estimate of the process that generated a given segment of the waveform corresponding to that state. Prior to speech recognition using HMMs, the waveform is transformed into a sequence of parameter vectors, each of which corresponds to some short segment of the waveform. Here, parameter vectors represent modeled auditory excitation patterns of the kind shown in Fig. 1.

The core process during recognition is a computation of the likelihood $f(x|C_i)$ that a HMM state C_i could have generated a parameter vector x. This process is repeated for each HMM state and each vector in the sequence, and the HMM that contains the most likely sequence of states is treated as the most likely model for the observed waveform. For the *continuous density* HMMs used in this study, the probability distribution is constrained to be a mixture of Gaussians,

$$f(x|C_1) = \sum_{k=1}^{M} P(k|C_i) f(x|k, C_i), \tag{1}$$

where the $P(k|C_i)$ are the mixture coefficients.

For missing-data ASR using continuous density HMMs, the likelihood computation step is modified as follows. The parameter vector x now consists of components that are regarded as clean and hence reliable, x_r, and other components that are masked, and hence unreliable, x_u. It can be shown (Cooke *et al.*, 2001) that, under a number of assumptions of the kind normally applied in ASR, the required likelihood is given by

$$f(x_r|C_i) = \sum_{k=1}^{M} P(k|C_i) f(x_r|k, C_i) \int f(x_u|k, C_i) dx_u. \tag{2}$$

If the unreliable components are missing altogether (i.e., no information about them is available, as might be the case if the signal had been filtered), then the integral in Eq. (2) reduces to unity, and the likelihood is evaluated as a weighted sum of partial likelihoods,

$$f(x_r|k, C_i) = N(x_r; \mu_{r,k,i}, \sigma^2_{r,k,i}), \tag{3}$$

where $N(x; \mu_{k,i}, \sigma^2_{k,i})$ denotes the Gaussian distribution for mixture component k in state i. Means, variances, and mixture coefficients are estimated during the training process. Equations (2) and (3) forms the basis for the *glimpses-only* model, described in Sec. IV B 3.

There are situations where the unreliable components do contain information. If the parameter vector x represents an energy-based quantity such as the modeled auditory excitation patterns used in this study, and if the unreliability is caused by masking from a noisy background, then the observed value can be treated as an upper bound, x_{high}, on the

energy of the masked speech component at that point in time and frequency. By taking some fixed lower bound for energy, x_{low}, it can be shown (Cooke *et al.*, 2001) that the integral in Eq. (2) is the difference of error functions:

$$\int f(x_u|k,C_i)dx_u = \frac{1}{2}\left[\text{erf}\left(\frac{x_{\text{high},u}-\mu_{u,k,i}}{\sqrt{2}\sigma_{u,k,i}}\right) - \text{erf}\left(\frac{x_{\text{low},u}-\mu_{u,k,i}}{\sqrt{2}\sigma_{u,k,i}}\right)\right]. \quad (4)$$

This computation forms the basis for the *glimpses-plus-background* model described in Sec. IV B 3.

III. SPEECH IN NOISE TASK

A. Speech and noise material

Speech tokens were drawn from the vowel-consonant-vowel (VCV) corpus collected by Shannon *et al.* (1999). The subset of 16 consonants /b, p, d, t, g, k, m, n, l, r, f, v, s, z, ʃ, tʃ/ in the vowel context /ɑ/ was employed. Of the 10 repetitions of each VCV from each talker,[I] 2 from each of 5 male talkers made up a test set of 160 items. The remaining 8 tokens of each VCV were used as a training set for the recognizer. Tokens were normalized to have equal rms energy.

Noise signals were formed by modulating speech-shaped noise with the envelope of *N*-talker babble for various *N*. Such signals produce approximately the same amount of energetic masking as the signal from which the envelope is drawn (Festen and Plomp, 1990) and were used here to limit informational masking effects. Following Brungart *et al.* (2001), the envelope was computed by convolving the absolute value of an *N*-talker babble signal with a 7.2 ms rectangular window. Babble was produced by summing utterances with equal rms energy from the TIMIT corpus (Garofolo *et al.*, 1992). Twelve babble-modulated noise conditions were employed corresponding to the following values of *N*: 1, 2, 3, 4, 6, 8, 16, 32, 64, 128, 512, and ∞ (i.e., speech-shaped noise).

Noisy tokens were produced by adding babble-modulated speech-shaped noise in each of the 12 noise conditions to the test part of the corpus at a global SNR of −6 dB, computed on a token-by-token basis. Since the energetic content of VCV syllables is dominated by the vowel portions, the actual SNR in the consonant portions was somewhat lower than −6 dB. Noise waveforms were gated on and off with the speech tokens.

B. Quantitative analysis of glimpsing opportunities

A quantitative analysis of glimpsing opportunities as a function of *N* was conducted. For this analysis, a glimpse was defined as a connected region of the spectrotemporal excitation pattern in which the energy of the speech token exceeded that of the background by at least 3 dB in each time–frequency element or "pixel." A "pixel" corresponds to a single time frame and frequency channel in the STEP representation. Here, elements of a region were deemed to be connected if they were part of the four-neighborhood (i.e., excluding diagonal neighbors) of any other element in the region. The "area" or extent of a glimpse was taken to be the number of time–frequency elements making up the glimpsed region. This is not an area in the traditional sense since the time and frequency units are not identical. Further, different choices of time and frequency resolution in the STEP will result in slight differences in calculated "areas." The choices here (defined in Sec. IV B 1) are based on those commonly used in studies employing STEPs.

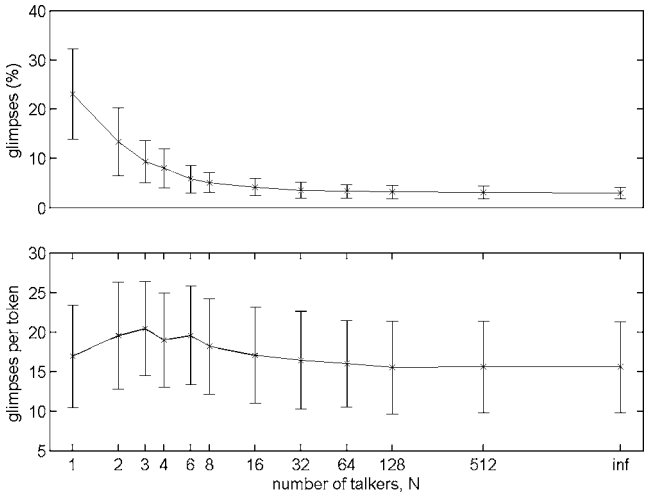

FIG. 2. Glimpse percentages and counts expressed as means across all tokens in the test set, separately computed for each noise condition. Error bars represent ±1 standard deviation. Here and elsewhere, the point marked "inf" indicates the speech-shaped noise condition.

Two quantities—glimpse area and glimpse count—were measured for each noise token in the corpus. Means and standard deviations of the two measures in each noise condition are shown in Fig. 2. The upper panel displays the mean percentage of each token covered by glimpses. While there is substantial variability across individual tokens, the mean glimpse percentage falls rapidly with *N*, leveling off at around 3% for $N>16$. The lower panel of Fig. 2 plots the number of glimpses per token. Interestingly, the range of means is quite small, suggesting that each noise condition results in a similar number of glimpses, although the quality of opportunities (as defined by the glimpse area) differs substantially as a function of *N*. These metrics confirm that qualitatively different glimpsing opportunities are available as the number of talkers contributing to the babble-modulated masker is varied, as demonstrated in Fig. 1.

C. Behavioral experiment

The intelligibility of consonants in babble-modulated noise was measured by Simpson and Cooke (2005) as part of a study comparing the masking effectiveness of natural babble and babble-modulated noise for varying numbers of talkers contributing to the babble. The babble-modulated noise conditions and speech tokens used by Simpson and Cooke (2005) were precisely the same as those used in the current study, as described in Sec. III A. Consonant identification scores based on 12 normal-hearing listeners are replotted in Fig. 3. In line with an earlier study that employed babble-modulated noise (Bronkhorst and Plomp, 1992), in-

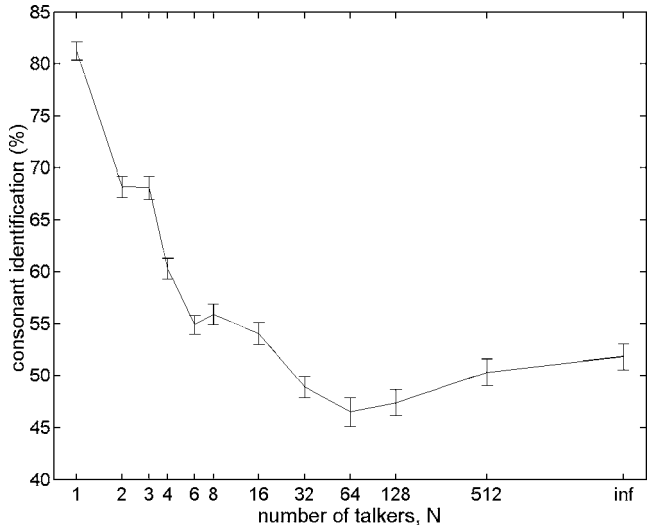

FIG. 3. Consonant intelligibility as a function of the number of talkers in the babble-modulated noise. Error bars represent ±1 standard error. Replotted from Fig. 1 of Simpson and Cooke (2005).

telligibility falls steeply with the number, N, of contributing talkers before leveling out for $N>32$. In fact, the intelligibility for all $N>6$ is not significantly different from speech-shaped noise.

Figure 4 plots the mean glimpse percentage in each noise condition against listeners' intelligibility results. The high correlation (0.955) between these measures suggests that the glimpse proportion alone is a very good linear predictor of intelligibility for these stimuli.

IV. GLIMPSING MODEL

A. Overview

Figure 5(a) illustrates the architecture of the glimpsing model for listeners. Glimpse detection is assumed to operate on a spectrotemporal representation produced in the early stages of auditory processing. Glimpses then have to be tracked through time and integrated to form a running speech

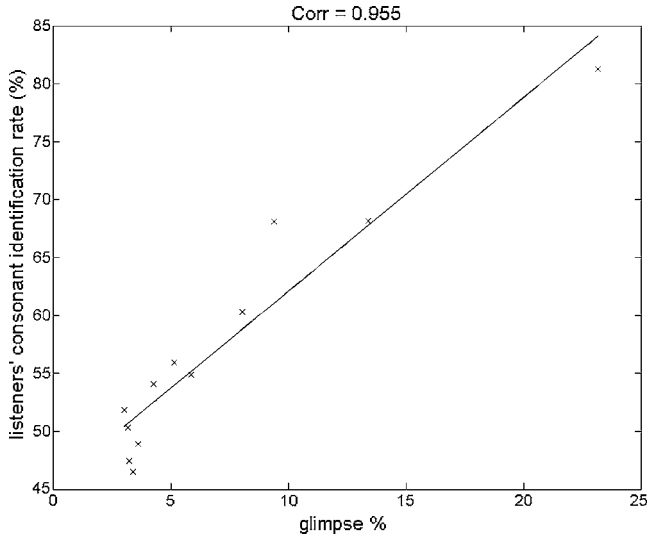

FIG. 4. The correlation between intelligibility and proportion of the target speech in which the local SNR exceeds 3 dB. Each point represents a noise condition, and proportions are means across all tokens in the test set. The best linear fit is also shown. The correlation between listeners and these putative glimpses is 0.955.

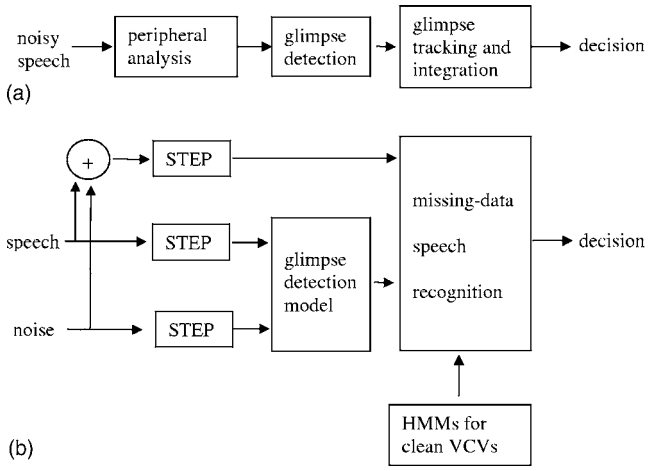

FIG. 5. (a) Architecture of the glimpsing model in humans. (b) Processing steps in the current study.

hypothesis. The temporal integration of glimpses may be based on estimates of pitch movements or speaker characteristics such as vocal tract length. Since the purpose of this study was to explore glimpse detection models that best account for behavioral findings rather than to provide a computational procedure for robust automatic speech recognition in noise, the actual processing steps used to test the model had access to both speech and noise signals [Fig. 5(b)]. STEPs for speech, noise, and their mixture were computed independently. The glimpse detection model determines those spectrotemporal regions dominated by speech. These regions constitute a mask that, along with the STEP for the mixture, was processed by a missing-data speech recognizer that computes the most likely speech hypothesis based on the partial evidence provided by the glimpses. The recognizer employed continuous density hidden Markov models derived by training on clean VCV tokens.

B. Methods

1. STEP computation

The spectrotemporal excitation pattern used as input to the model is a smoothed and compressed representation of the envelope of the basilar membrane response to sound. The waveform was initially processed by a bank of 40 gammatone filters (Patterson *et al.*, 1988) implemented using a pole-mapping procedure described in Cooke (1993). Filters were equally spaced on an ERB-rate scale between 50 and 7500 Hz. The Hilbert envelope in each channel of the filterbank was computed and smoothed with a leaky integrator with an 8 ms time constant (Moore *et al.*, 1988). The smoothed envelope was downsampled to 100 Hz and log-compressed.

2. Glimpse detection

One of the purposes of the current study was to explore the effect of different assumptions about what constitutes a usable and detectable glimpse. The glimpses shown in Fig. 1 result from a detection model which assumes that all spectrotemporal elements whose local SNR exceeds 3 dB are detectable and used in a subsequent classification of the pat-

tern. However, there is no prior reason for choosing this particular local SNR threshold. Further, the assumption that listeners can detect very small regions of favorable local SNR when surrounded by a masker may be unreasonable. The recognition experiments reported here treated both local SNR for the detection and minimum glimpse area as free parameters over which the match to listeners' performance was optimized. The output of the glimpse detection stage is a time–frequency mask.

3. Missing-data recognition

The glimpse mask and STEP representation of the mixture signal provide the input to a missing-data speech recognizer. This study employed the HMM-based missing-data methods described in detail by Cooke *et al.* (2001) and reviewed in Sec. II. Two variants of the missing-data technique were used. The first, referred to as the glimpses-only model, employed information in the glimpsed regions alone and ignored information in the masked regions. The second, the glimpses-plus-background model, used, in addition, information from the masked regions. The glimpses-plus-background model was motivated by the phoneme restoration effect (Warren, 1970) and the subsequent demonstration of spectral induction (Warren *et al.*, 1997). These studies used speech with missing temporal fragments or spectral regions, respectively. In both cases, listeners were more likely to perceive or identify sounds if the missing parts were replaced by noise with sufficient energy to have masked the missing portion, had it been present.

The missing-data recognizer used a continuous density HMM for each of the 16 VCVs. Each VCV model had six emitting states, while each state employed a four-component Gaussian mixture density. Here, 40 examples of each VCV (8 from each of the 5 male talkers, distinct from those used for evaluation) were used to train the HMMs. Model training was performed using the HMM toolkit HTK (Young and Woodland, 1993) while recognition used a local implementation of a missing-data recognizer. Performance in the no-noise condition exceeded 99%. For the glimpses-plus-background model, a lower energy bound of −120 dB was used for x_{low} in Eq. (4).

C. Results

1. Variation in local SNR threshold for glimpse detection

Figure 6 compares listeners' and model recognition rates in each noise-masking condition, expressed as the percentage of consonants correctly identified across the test set, for a single local SNR detection threshold. The closeness of the match obtained for this particular choice of detection model suggests that glimpsing can account for the gross pattern of intelligibility of VCVs in stationary and nonstationary backgrounds.

Figure 7 compares listeners' and model recognition rates as a function of the local SNR for glimpse detection. The upper panel plots the root-mean-square (rms) listener–model distance computed across noise-masking conditions (e.g., the rms distance between the two curves shown in Fig. 6). This distance is a positive quantity and measures the proximity of

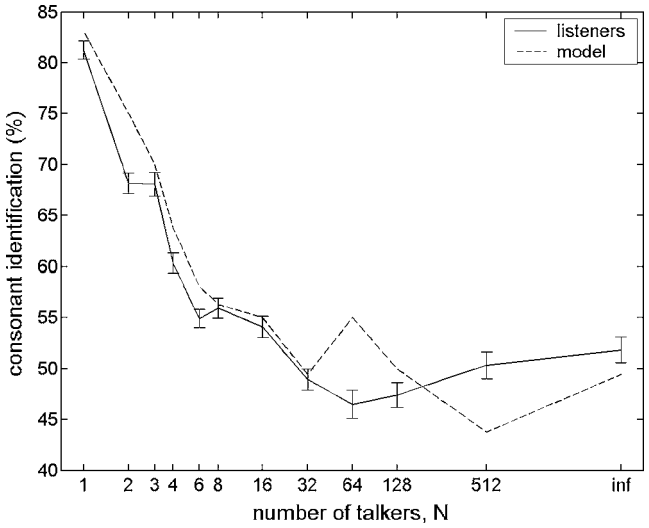

FIG. 6. Model and listeners' consonant identification rates for the glimpses-only model at a local SNR detection threshold of 6 dB. Listeners' data here and in Fig. 9 are the same as that of Fig. 3.

the two curves. The lower panel plots the mean listener-model difference or "bias." This is a signed quantity that reflects the extent to which the model outperforms listeners (positive bias values) or *vice versa* (negative bias values).

Both the glimpses-only and glimpse-plus-background models result in a pair of local minima in the model-listener rms distance. A consideration of the effect of changing the local SNR for detection explains the presence of two minima. At increasingly large negative local SNRs, many regions are treated as glimpses. However, some of these will be dominated by the background source. Consequently, model performance is dominated by errors due to distortions in the glimpsed data and its performance falls below that of listeners, as shown by the bias curve in the lower panel of Fig. 7. At the other extreme of large positive local SNRs for detection, glimpse scarcity limits performance since the number of regions satisfying increasingly strict criteria for inclusion falls as the local SNR increases. Again, the bias curve indicates that model performance suffers relative to listeners in such conditions.

Between the two extremes of glimpse corruption and glimpse scarcity, there is a broad region (from −2 to +6 dB for the glimpses-only model and from −5 to +6 dB for the glimpses-plus-background model), where the model outperforms listeners. Local minima in the listener–model distance occur at the edges of the range. This result suggests that there is more than sufficient information in glimpses to support their use as the basis for speech perception in noise, but that listeners do not possess an optimal glimpse detection model.

The broad pattern of results is similar for both glimpses-only and glimpses-plus-background recognition variants. However, the curves differ in several details. The glimpses-plus-background model outperforms the glimpses-only model, as shown by the more extensive range of SNRs with positive bias in Fig. 7, reflecting the additional information used in the classification process. The closest listener-model match at negative SNR thresholds are of a similar size in the two models, but occur at different points (−2 vs −5 dB), perhaps because the additional information employed by the

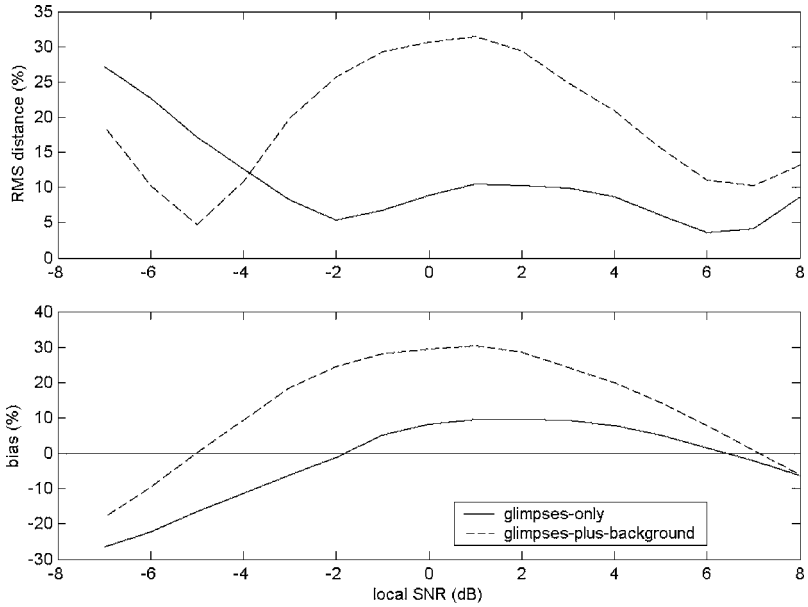

FIG. 7. The rms distance (upper panel) and mean difference (lower panel) between the listeners and model performance across all noise conditions as a function of the local SNR threshold for glimpse detection for the glimpses-only and glimpses-plus-background models.

glimpses-plus-background model allows it to tolerate more distortion in the evidence considered to be part of the speech hypothesis.

2. Variation in glimpse area for detection

Since the basic glimpse detection model outperforms listeners across a broad range of local SNRs for detection, a number of more sophisticated detection models that further reduce the set of glimpses could also account for the results. Based on the assumption that listeners may be unable to detect very brief regions of target dominance, or regions that occupy a very narrow portion of the spectrum, a minimum glimpse area criterion was incorporated into the glimpse detection model. Specifically, all connected regions of spectrotemporal elements satisfying a given local SNR criterion also had to possess an "area" (as defined in Sec. III B) greater than a specified amount.

Figure 8 plots the rms model-listener distance as a function of the local SNR threshold and minimum glimpse area. An inspection of the bias functions illustrates that when glimpses with a minimum area of 25 were removed, the glimpses-only model underperformed listeners, while the glimpse-plus-background variant tolerated removal of glimpses with 75–100 elements before performance fell be-

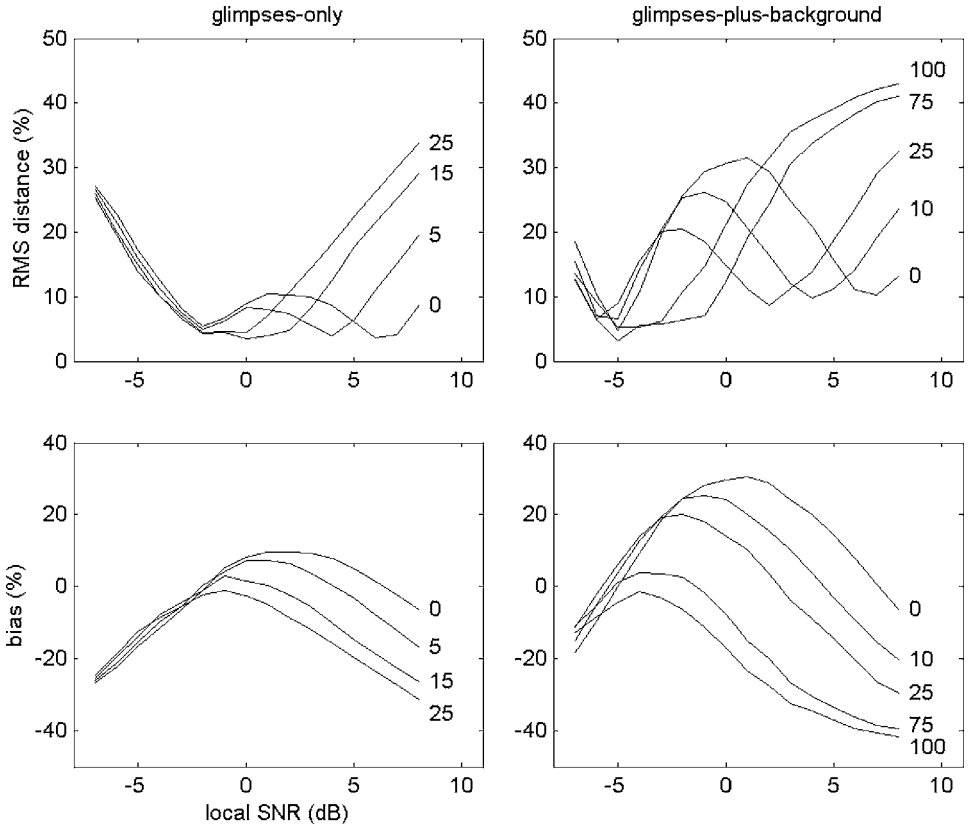

FIG. 8. The rms distance and bias for glimpse detection models in which both the local SNR for detection and the minimum glimpse area are varied. Numbers attached to curves identify minimum glimpse areas used.

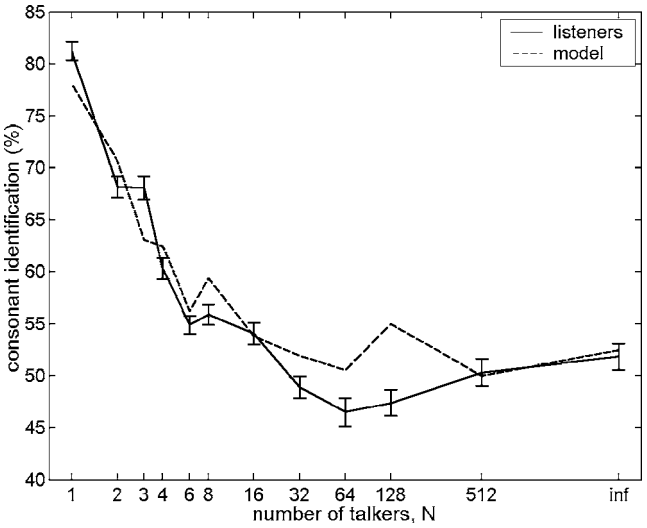

FIG. 9. Best listener-model fit (glimpses+background variant, local SNR for detection=−5 dB, minimum glimpse areas=75).

low that of listeners. The superior robustness of the latter model demonstrates that masked regions contain useful information for the discrimination between hypotheses about the identity of the speech target, in spite of the fact that no energy separation is performed in such regions. A striking feature of the results is the sensitivity of the best-fit location at positive local SNRs. As increasingly larger glimpses are excluded, the position of the minimum rapidly shifts to lower local SNRs, presumably because of the decreasing amount of evidence on which to base identification. In contrast, the location of the other local minimum barely changes as data are excluded. This finding suggests that the best fit at a negative glimpse detection threshold is robust, even though some glimpses are corrupted by energy from the background. The best fit occurred at a local SNR for detection of −5 dB within a minimum glimpse area of 75 elements for the glimpses-plus-background model (Fig. 9). All further analyses were based on this best-fitting model.

3. Consonant identification rates and confusions

Figure 10 compares listeners' and model identification rates for each consonant, averaged across noise conditions. While the overall identification rates are very similar (listeners: 57.7%, model: 57.5%), there are clear differences in scores for individual consonants. The model always outperformed listeners in the identification of plosives and nonsibilant fricatives (/v/ and /f/). The difference in performance was sometimes large (e.g., for /b/ and /f/). Listeners scored more highly than the model for the sibilants /s/, /ʃ/, and /tʃ/ and for the nasals. In the case of /m/, the margin was large. The source of these disparities is not clear.

Figure 11 depicts model and listener identification rates for each consonant as a function of noise condition. These results show that the overall fall in intelligibility as *N* increases is not always reflected in the recognition rates of individual consonants. Recognition of the plosives (/t/ and /g/) and the affricate /tʃ/ was uniformly good across differing masking conditions, while /b/ and /v/ were difficult to identify in the presence of most maskers. The distribution of available cues in a noise modulated by the envelope of a single talker will be quite different from those that remain after masking by stationary noise, and it may be that the multiple, redundant cues to identity are better distributed in some consonants than in others.

Tables I and II report consonant confusions for listeners and the best-fitting model, respectively. At the detailed level of individual confusions, it is clear that the model differs from listeners. However, some common trends can be seen. Both model and listeners tend to over-report /t/. Similarly, /m/ and /v/ tokens result in a wide range of response alternatives. Table III shows the proportion of transmitted information for voicing, manner and place for listeners and the best-fitting model (Miller and Nicely, 1955; Wang and Bilger, 1973). This analysis reveals that while manner and place information is transmitted at a similar rate in the listeners and model, voicing information is far more susceptible to masking in the latter. This may be a defect of the excitation pattern representation employed by the model. Implicit cues to voicing are conveyed by the resolved harmonics in the lower-frequency region of the excitation pattern, but temporal cues that would indicate the presence of voicing are lost. Spectral averaging in the formation of HMM probability densities may also contribute to the degradation of voicing information (Barker, 1998).

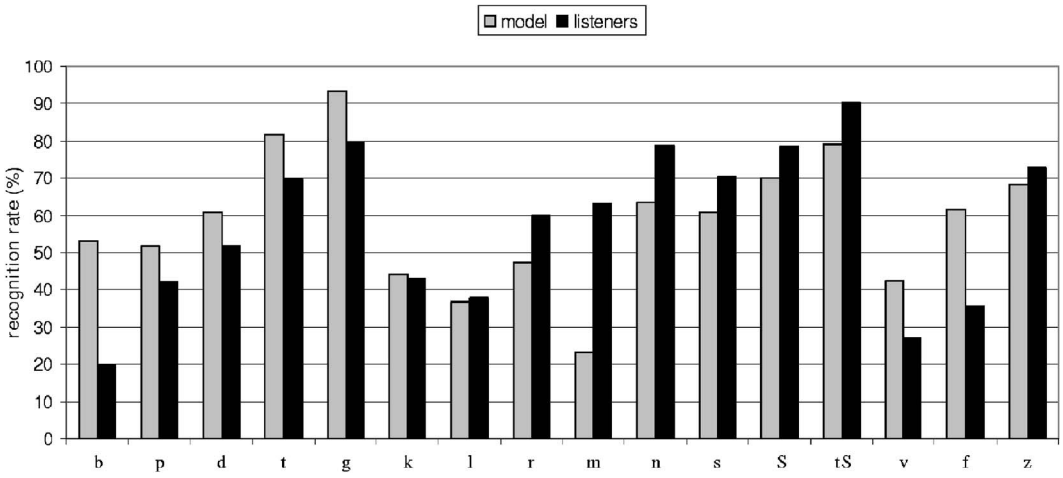

FIG. 10. Consonant identification scores for listeners and the best-fitting model.

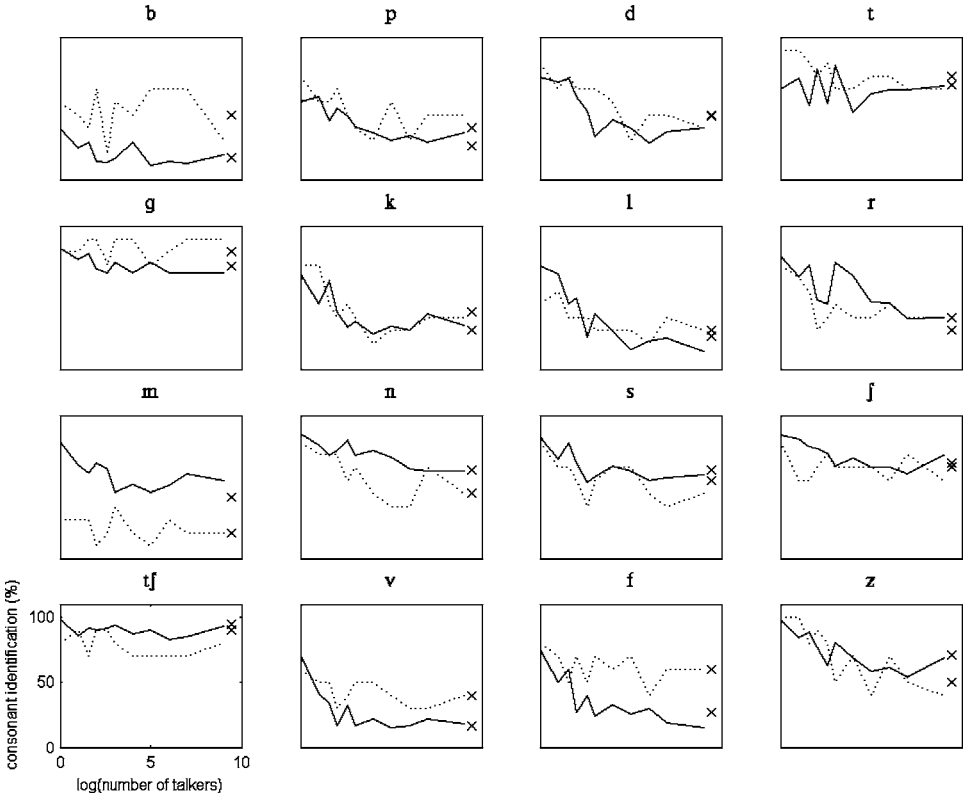

FIG. 11. Consonant identification scores as a function of N (solid line: listeners, dotted line: model). Crosses depict scores in the speech-shaped noise condition.

4. Response set probability

A valid model of speech perception should be able to predict the most likely judgements made by a set of listeners to each individual noisy token in an evaluation set. Such an evaluation criterion is extremely demanding, given our current understanding of speech perception in noise. A suitable metric was developed and evaluated on the current task. The metric computes the log probability of the entire response sequence. The response probability for an individual token can be computed from the discrete probability distribution formed from the set of listener responses to that token. This method requires a sufficiently large number of judges in order to provide a robust estimate of the response probability distribution. The probability of the entire sequence—the response set probability—is the product of the individual response probabilities.

Response set probabilities for each listener and the model can be compared. The probability for any individual listener is obtained by leaving out their judgements and using the judgements of the other listeners to form the probability distribution for each token. Statistical hypothesis testing can

TABLE I. Confusion matrix for listeners, averaged over noise conditions. Numbers represent percentage responses. For clarity, response percentages smaller than 1.5% have been removed. Consequently, column totals may not match the sum of column entries. Confusions greater than or equal to 10% are presented in bold. Each (row, column) entry signifies the (stimulus, response) confusion.

	b	p	d	t	g	k	l	r	m	n	s	ʃ	tʃ	v	f	z
b	**21**	9	5	3	4	4	6	5	7	3				**21**	9	2
p	3	**43**		7		**12**	2	2	7	2				6	**12**	
d			**52**	4	**18**	2	3	2		**11**			2	2		3
t		6	4	**70**		9			2	3					2	
g	2		3		**80**	3		3						2		
k		**14**	2	**10**	6	**43**			4	2			2	2	**10**	
l	5	5	2	2	3		**37**	8	**20**	2				**9**	4	
r	4	4		2	3		5	**59**	9					7	4	
m	3	4		2	2	2	7	4	**63**	7				3		
n			5	3	3		3		4	**79**						
s		2		2		2					**71**				9	9
ʃ												**79**	**15**			
tʃ				2								4	**90**			
v	**12**	7	3	3	4	3	7	7	**14**	2				**27**	6	4
f	6	**23**		4	2	9		2	8					8	**35**	
z			8	3	2	2	2			2			2	2		**73**

TABLE II. Confusion matrix for the best-fitting model.

	b	p	d	t	g	k	l	r	m	n	s	ʃ	tʃ	v	f	z
b	**53**		**12**			**11**	3						6	8	5	
p	3	**52**		**17**			6	4					8	8		3
d	5		**61**	9	8	4							8	3		
t		8	3	**82**		3							3			2
g			3	3	**93**											
k	9	6	3	3	3	**44**		4			3		**13**	3	8	
l	**10**		3	7		**21**	**37**	**17**						3	2	
r	6	3	6	4	2	2	**10**	**48**			**11**			4	2	3
m		**13**	2	**18**		2	4	6	**23**	7	3	3	6	7		8
n				**16**	**10**	2			2	**63**				4		2
s				**22**							**61**					**18**
ʃ												**70**	**27**	2		
tʃ												**18**	**79**			
v	3	6	4	8		9	**10**				4		9	**43**	2	2
f	4		8	5							5	2		**13**	**62**	
z	5	3		7									**10**	5		**68**

TABLE III. Proportion of transmitted information for voicing, manner, and place. Consonant subsets used for each distinction are indicated.

	Listeners	Model
Voicing [b d g l r m n v z] [p t k s ʃ tʃ f]	0.424	0.292
Manner [b d g p t k] [tʃ] [l r] [m n] [v z s ʃ f]	0.397	0.351
Place [b p m f v] [d t n s l z] [g k] [r ʃ tʃ]	0.371	0.367

then be used to determine the probability that the model belongs to the listeners' distribution of response likelihoods.

Since the current study used a relatively small number of listeners, it is difficult to obtain robust estimates of per token response probabilities. One specific issue is the problem of zero probabilities for certain response categories. For instance, in response to an /s/ token in noise, it may be that no listener in the set responded with /g/, leading to a zero probability for that response. If the model were to respond with /g/, the probability of the entire response set becomes zero. One solution is to replace zero probabilities with a certain minimum value and renormalize the distribution.

Response set probabilities for each noise condition were computed for all listeners and for the best-fitting model. A probability floor of 0.1 was used. Log response set probabilities are plotted in Fig. 12. They show that, in all masking conditions, listeners act as a cohort and are more similar to each other than to the model. Listeners' responses are most similar to each other in masking conditions, which lead to a high overall performance due to the lower incidence of confusions.

V. GENERAL DISCUSSION

The main finding of this study is that glimpses contain more than enough information to support speech recognition in a computational model. This suggests that a glimpsing process may serve as a basis for human speech perception in noise. Several glimpsing models can account for listeners' performance on a consonant in noise task. A model that used solely the information in the putative glimpse regions produced a close fit to the behavioral data by assuming that (i) glimpses are detected if the local SNR is greater than 0 dB, and (ii) regions of positive local SNR that do not possess a sufficient spectrotemporal extent will not be detected. A model that additionally exploited information in the masked regions also matched listeners', data but with quantitatively different assumptions. A robust feature of the latter model was the requirement that glimpses be detectable when the local SNR is greater than −5 dB.

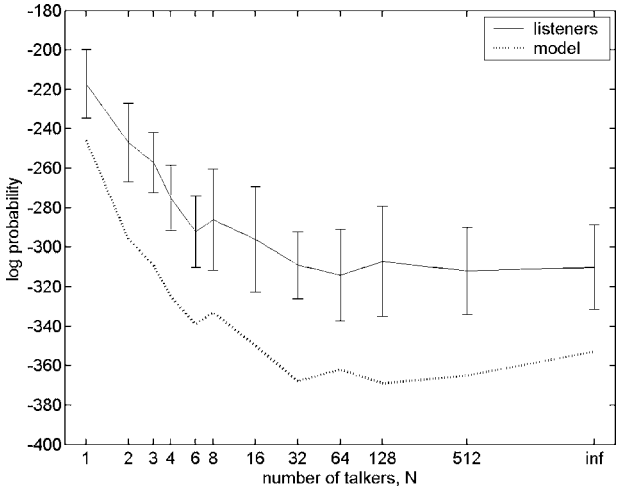

FIG. 12. Log probability of the response set as a function of *N*. Listeners: solid line, model: dotted line. Error bars denote ±2 standard deviations.

The optimum local SNR detection thresholds for the glimpses-only and glimpses-plus-background models suggest somewhat different solutions to the subsequent identification problem. A 0 dB threshold produces fewer glimpses, but those that are available are reasonably undistorted. At a detection threshold of −5 dB, many more regions will be considered as glimpses, but some of them will be significantly contaminated by background energy. The latter value is consistent with the model of Moore *et al.* (1997), which predicts a threshold of −4 dB for detecting complex signals in noise, and with the findings of Brungart *et al.* (submitted), who suggested an optimum local SNR threshold of −6 dB. However, the study of Drullman (1995) found that speech components more than 2 dB below the noise level in a given auditory filter made no contribution to intelligibility.

The glimpse detection models in the current study used simulated spectrotemporal excitation patterns that are good first-order representations of auditory stimuli at an early stage of processing. However, the excitation pattern representations used here are deficient in two respects. They lack both a representation of the fine structure of auditory-nerve fiber response to sound and a model of nonsimultaneous masking. It seems likely that a more realistic model of the temporal fine structure would result in glimpses differing in detail from those in the model presented here. However, the changing local dominance relations between the target and background that give rise to glimpsing opportunities are likely to affect other representations in broadly similar ways. For instance, the temporal response characteristic seen in auditory-nerve fibers discharge patterns to locally strong stimulus components known as "synchrony capture" can be considered as a reflection of the stimulus dominance at the level of a temporal fine structure (e.g., Sinex and Geisler, 1983).

If speech perception in noise is based on glimpsing, listeners must have developed solutions to two fundamental problems—detection and integration. The detection models described here make use of foreknowledge of both local SNR and glimpse extent. Several computational attempts to estimate local SNR have been reported. Berthommier and Glotin (1999) showed that the disruption to harmonicity was a good predictor of local SNR, while Tchorz and Kollmeier (2002) estimated local SNR from amplitude modulation spectrograms using a neural network learning procedure.

Rather than estimating local SNR directly, it is possible that listeners apply principles of auditory organization to group spectrotemporal regions based on properties such as the similarity of location or fundamental frequency (Bregman, 1990; Darwin and Carlyon, 1995; Cooke and Ellis, 2001; Brown and Wang, 2005). An algorithm that could be used for glimpse estimation based on binaural cues was pre-

sented in Roman *et al.* (2003). Alternatively, it might be possible to identify clean fragments using prior speech models.

The second problem is to integrate glimpses into an ongoing speech hypothesis. Assmann and Summerfield (2004) call this the tracking problem. While integration may not represent a significant challenge for a single speech target in modulated babble noise since there is little possibility of confusion between this type of noise and speech, it becomes more important when the background contains other talkers. In such conditions, listeners have to contend with both energetic and informational masking (Carhart *et al.*, 1969; Brungart *et al.*, 2001; Freyman *et al.*, 2004). In the glimpsing account, informational masking can be understood as arising from the incorrect assignment of glimpses to the developing speech hypothesis. Glimpse integration may be less of a problem in practice than detection. A recent computational approach to glimpse integration (Barker *et al.*, 2005) uses prior speech models to track fragments of evidence through time.

The approach adopted here belongs to a small class of models of human speech recognition that exploit speech models and automatic speech recognition techniques (Ainsworth and Meyer, 1994; Ghitza, 1993; Holube and Kollmeier, 1996). ASR provides a well-founded basis for constructing models of speech schema, although it requires the use of larger corpora than are typically employed in perceptual studies. Direct computational models contrast with macroscopic models such as the articulation index (Kryter, 1962), speech transmission index (Steeneken and Houtgast, 1979), and speech intelligibility index (ANSI, 1997), which attempt to predict speech intelligibility in conditions of background noise, reverberation, and channel distortion. In principle, direct computational models can also be applied to speech intelligibility assessment, and promise to provide more detailed information in a wider range of conditions than those handled by macroscopic approaches. For instance, a direct computational model employing a sophisticated statistical model of speech can produce decisions for individual tokens in arbitrarily complex noise backgrounds. However, representations of speech and associated decoding processes used in ASR are likely to be rather different from those employed by listeners. It is clear that current ASR-based models of human speech perception are incapable of predicting listeners' responses in detail. Here, a response set probability metric indicated that the panel of listeners were more similar to each other than to the model, in spite of the existence of a close model-listener match at the level of overall performance. At the level of consonant confusions, clear differences between the current model and behavioral data remain. Information theoretic analyses suggested that the model is weak on the transmission of voicing cues. Whether this represents a deficiency in the spectrotemporal excitation pattern representation itself or stems from a loss of information during the formation of speech schema from individual training tokens is a matter for further study.

VI. CONCLUSIONS

A processing model that takes advantages of relatively clear views of a target speech signal can account for the overall intelligibility of VCV tokens in a range of energetic masking conditions. As a strategy for dealing with speech in noise for listeners and algorithms, focusing on the regions with advantageous local SNR may be simpler than estimating the energy contribution of the speech signal at each point in time and frequency. The best fit to behavioral data came from a model that (i) used information in the glimpses and counterevidence in the masked regions; (ii) restricted glimpses to have a certain minimum area; and (iii) treated all regions with local SNR in excess of −5 dB as potential glimpses. However, many different glimpse detection strategies have the capacity to explain the observed level of listeners' performance.

Attempts to use techniques from automatic speech recognition to model human speech perception are in their infancy, but this study suggests that insights into both ASR and HSR can be gained by a modeling approach that tests different potential perceptual strategies for handling speech in noise.

ACKNOWLEDGMENTS

This work was supported by Grant No. GR/R47400/01 from the UK Engineering and Physical Science Research Council. The author would like to thank Maria Luisa Garcia Lecumberri, Jon Barker, and Sarah Simpson for valuable comments on the manuscript, and Professor Bob Shannon for making available the extended version of the VCV corpus. Three anonymous reviewers provided helpful comments.

[1]The published corpus released a single example of each VCV from every talker, but ten were recorded. Professor Shannon kindly made the original recordings available to the author.

Ainsworth, W. A., and Meyer, G. F. (**1994**). "Recognition of plosive syllables in noise: Comparison of an auditory model with human performance," J. Acoust. Soc. Am. **96**, 687–694.

Alcantara, J. I., Weisblatt, E. J. L., Moore, B. C. J., and Bolton, P. F. (**2004**). "Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome," J. Child Psychol. Psychiatry **45**, 1107–1114.

ANSI (1997). American National Standard Methods for Calculation of the Speech Intelligibility Index Document Number: ANSI/ASA S3.5-1997.

Assmann, P. F. (**1996**). "Tracking and glimpsing speech in noise: role of fundamental frequency," J. Acoust. Soc. Am. **100**, 2680.

Assmann, P. F., and Summerfield, Q. (**1994**). "The contribution of waveform interactions to the perception of concurrent vowels," J. Acoust. Soc. Am. **95**, 471–484.

Assmann, P. F., and Summerfield, Q. (**2004**). "The perception of speech under adverse acoustic conditions," in *Speech Processing in the Auditory System*, Springer Handbook of Auditory Research, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer-Verlag, Berlin), Vol. 18.

Barker, J. P. (**1998**). "The relationship between auditory organisation and speech perception: Studies with spectrally reduced speech," unpublished Ph.D. thesis, University of Sheffield.

Barker, J., and Cooke, M. P. (**1997**). "Modelling the recognition of spectrally reduced speech," *Proc. EUROSPEECH*, pp. 2127–2130.

Barker, J., Cooke, M. P., and Ellis, D. P. W. (**2005**). "Decoding speech in the presence of other sources," Speech Commun. **45**, 5–25.

Berthommier, F., and Glotin, H. (**1999**). "A new SNR feature mapping for robust multistream speech recognition," *Proc. XIVth Int. Cong. Phonetic Sciences*, pp. 711–715.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT Press, Cambridge, MA).
Bronkhorst, A. W., and Plomp, R. (**1992**). "Effect of multiple speech-like maskers on binaural speech recognition in normal and impaired hearing," J. Acoust. Soc. Am. **92**, 3132–3139.
Brown, G. J., and Wang, D. (**2005**). "Separation of speech by computational auditory scene analysis, in *Speech Enhancement*, edited by J. Benesty, S. Makino and J. Chen (Springer-Verlag, New York), pp. 371–402.
Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **100**, 2527–2538.
Brungart, D. S., Chang, P., Simpson, B. D., and Wang, D. (submitted). "Isolating the energetic component of speech-on-speech masking with an ideal binary time–frequency mask," J. Acoust. Soc. Am..
Buss, E., Hall, J. W., and Grose, J. H. (**2003**). "Effect of amplitude modulation coherence for masked speech signals filtered into narrow bands," J. Acoust. Soc. Am. **113**, 462–467.
Buss, E., Hall, J. W., and Grose, J. H. (**2004**). "Spectral integration of synchronous and asynchronous cues to consonant identification," J. Acoust. Soc. Am. **115**, 2278–2285.
Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**, 694–703.
Comon, P. (**1994**). "Independent component analysis. A new concept?," Signal Process. **36**, 287–314.
Cooke, M. P. (**1993**). *Modelling Auditory Processing and Organisation* (Cambridge University Press, Cambridge).
Cooke, M. P. (**2003**). "Glimpsing speech," J. Phonetics **31**, 579–584.
Cooke, M. P., and Ellis, D. P. W. (**2001**). "The auditory organization of speech and other sources in listeners and computational models," Speech Commun. **35**, 141–177.
Cooke, M. P., Green, P. D., and Crawford, M. D. (**1994**). "Handling missing data in speech recognition," *Proc. 3rd Int. Conf. Spoken Language Processing*, 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.
Culling, J. F., and Darwin, C. J. (**1994**). "Perceptual and computational separation of simultaneous vowels: cues arising from low frequency beating," J. Acoust. Soc. Am. **95**, 1559–1569.
Cunningham, S. P. (**2003**). "Modelling the recognition of band-pass filtered speech," Doctoral thesis, Department of Computer Science, The University of Sheffield.
Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *The Handbook of Perception and Cognition, Vol. 6, Hearing*, edited by B. C. J. Moore (Academic, New York), pp. 387–424.
de Cheveigné, A., and Kawahara, H. (**1999**). "Missing-data model of vowel identification," J. Acoust. Soc. Am. **105**, 3497–3508.
Drullman, R. (**1995**). "Speech intelligibility in noise: relative contributions of speech elements above and below the noise level," J. Acoust. Soc. Am. **98**, 1796–1798.
Drygajlo, A., and El-Maliki, M. (**1998**). "Speaker verification in noisy environment with combined spectral subtraction and missing data theory," *Proc. ICASSP-98*, pp. 121–124.
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.
Fletcher, H. (**1953**). *Speech and Hearing in Communication* (Van Nostrand, New York).
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.
Gales, M. J. F., and Young, S. J. (**1993**). "HMM recognition in noise using parallel model combination," *Proc. EUROSPEECH*, pp. 837–840.
Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1992**). "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," NIST.
Ghitza, O. (**1993**). "Adequacy of auditory models to predict human internal representation of speech sounds," J. Acoust. Soc. Am. **93**, 2160–2171.
Gustafsson, H. A., and Arlinger, S. D. (**1994**). "Masking of speech by amplitude-modulated noise," J. Acoust. Soc. Am. **95**, 518–529.
Holube, I., and Kollmeier, B. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1716.
Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**, 2915–2922.
Hyvärinen, A., Karhunen, J., and Oja, E. (**2001**). *Independent Component Analysis* (Wiley, New York).
Kasturi, K., Loizou, P. C., Dorman, M., and Spahr, T. (**2002**). "The intelligibility of speech with 'holes' in the spectrum," J. Acoust. Soc. Am. **112**, 1102–1111.
Kryter, K. D. (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.
Lippmann, R. P. (**1996**). "Accurate consonant perception without mid-frequency speech energy," IEEE Trans. Speech Audio Process. **4**, 66–69.
Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.
Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.
Miller, G. A., and Nicely, P. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Moore, B. C. J. (**2003**). "Temporal integration and context effects in hearing," J. Phonetics **31**, 563–574.
Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness, and partial loudness," J. Audio Eng. Soc. **45**, 224–240.
Moore, B. C. J., Glasberg, B. R., Plack, C. J., and Biswas, A. K. (**1988**). "The shape of the ear's temporal window," J. Acoust. Soc. Am. **83**, 1102–1116.
Palomäki, K. J., Brown, G. J., and Barker, J. (**2002**). "Missing data speech recognition in reverberant conditions," *Proc. ICASSP*, pp. 65–68.
Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1988**). "SVOS Final Report: The Auditory Filterbank," Technical Report 2341, MRC Applied Psychology Unit.
Raj, B., Singh, R., and Stern, M. (**1998**). "Inference of missing spectrographic features for robust speech recognition," *Proc. ICSLP*, pp. 1491–1494.
Roman, N., Wang, D., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.
Scott, S. K., Rosen, S., Wickham, L., and Wise, R. J. S. (**2004**). "A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception," J. Acoust. Soc. Am. **115**, 813–821.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.
Simpson, S., and Cooke, M. P. (**2005**). "Consonant identification in *N*-talker babble is a nonmonotonic function of *N*," J. Acoust. Soc. Am. **118**, 2775–2778.
Sinex, D. G., and Geisler, C. D. (**1983**). "Responses of auditory-nerve fibres to consonant-vowel syllables," J. Acoust. Soc. Am. **73**, 602–615.
Steeneken, H. J. M., and Houtgast, T. (**1979**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.
Tchorz, J., and Kollmeier, B. (**2002**). "Estimation of the signal-to-noise ratio with amplitude modulation spectrograms," Speech Commun. **38**, 1–13.
Varga, A. P., and Moore, R. K. (**1990**). "Hidden Markov model decomposition of speech and noise," *Proc. ICASSP*, pp. 845–848.
Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: a study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.
Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.
Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (**1995**). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," Percept. Psychophys. **57**, 175–182.
Warren, R. M., Hainsworth, K. R., Brubaker, B. S., Bashford, J. A., and Healy, E. W. (**1997**). "Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps," Percept. Psychophys. **59**, 275–283.
Young, S. J., and Woodland, P. C. (**1993**). "HTK Version 1.5: User, Reference and Programmer Manual," Cambridge University Engineering Department, Speech Group.