Modelling speaker intelligibility in noise

Jon Barker *, Martin Cooke

University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello Street, Sheffield, S1 4DP

Abstract

This study compared behavioural performance on a multispeaker speech-in-noise task with that of a model inspired by automatic speech recognition techniques. Listeners identified 3 keywords in simple 6-word sentences in speech-shaped noise spoken by one of 18 male or 16 female speakers. An across-speaker analysis of a number of acoustic parameters (vocal tract length, mean fundamental frequency and speaking rate) found none to be consistently good predictors of relative intelligibility. A simple measure of degree of energetic masking was a good predictor of female speech intelligibility, especially in high noise conditions, but failed to account for interspeaker differences for the male group. A glimpsing model, which combined a simulation of energetic masking with speaker-dependent statistical models, produced recognition scores which were fitted to the behavioural data pooled across all speakers. Using a single set of speaker-independent, noise-level-indepedent parameters, the model was able to predict not only the intelligibility of individual speakers to a remarkable degree, but could also account for most of the token-wise intelligibilities of the letter keywords. The fit was particularly good in high noise conditions.

1 Introduction

It is common experience that in noisy situations some speakers are consistently more intelligible than others. A large degree of variation in inter-speaker intelligibility persists even if the signal to noise ratio (SNR) is similar across speakers. The joint behavioural-modelling study described in this paper attempts to account for relative speaker intelligibility of utterances mixed with

* Corresponding author.
Email addresses: j.barker@dcs.shef.ac.uk (Jon Barker), m.cooke@dcs.shef.ac.uk (Martin Cooke).

stationary noise by combining models of energetic masking and speaker acoustics.

Previous studies of speaker intelligibility have focused on clean speech and typically examined a range of speakers, comparing those speakers that are 'intrinsically clear' with those that are less intelligible. Such studies attempt to identify a small set of acoustic parameters which can best predict relative speaker intelligibility. For example, in a study of intrinsically clear speech, Bond and Moore (1994) found intelligibility to be related to duration cues, the geometry of the vowel space and cues to consonant discrimination. Bradlow et al. (1996) also demonstrated an effect vowel space size and an additional effect of F0 range. In a more recent study involving a larger number of speakers, Hazan and Markham (2004) found that word intelligibility was significantly correlated with word duration and energy in the 1 to 3 kHz region, and that female speakers had a higher intelligibility than male speakers. However, if intelligibility is related to acoustic-phonetic characteristics, the relation is not simple – previous studies have all found great variability in the profile of the clearest speakers.

Another body of work contrasts casual speech with speech produced in a deliberately clear manner ('clear speech', Picheny et al., 1985). The most consistent clear speech effect is a reduction in speech rate (Picheny et al., 1985, 1986). Krause and Braida (2004) controlled for speaking rate by using speakers who could produce clear speech at normal speaking rates and found that that clear speech had increased energy in the 1 to 3 kHz range. At least in this respect, deliberately clear speech is similar to intrinsically clear speech.

The current work differs from these earlier studies in that it focuses on speech in noise. Noise introduces masking and the resulting intelligibility will be influenced by both intrinsic intelligibility and the speech-masker relationship. Intrinsically clear speakers are not necessarily the most intelligible in noisy conditions. Indeed, characteristics that produce noise robustness might have an adverse effect on intrinsic intelligibility. Most previous studies of speech in noise have examined the largely involuntary way in which speakers adapt their speaking style in adverse environments, the so-called Lombard effect. Acoustic analyses made by van Summers et al. (1988) and Junqua (1993) have shown that there is an increase in amplitude, changes to formant frequency and bandwidth, a change in spectral tilt and an increase in duration. Lombard speech differs acoustically from deliberately clear speech (van Summers et al., 1988). Our study does not consider the Lombard effect, but instead artificially mixes noise with speech recorded in clean conditions.

Noise reduces the intelligibility of speech due to two types of masking – 'energetic' and 'informational'. *Energetic masking* occurs in the periphery of the auditory system when the noise energy is greater than the speech energy in

some spectro-temporal region. The loss of information may reduce the discriminability between speech classes by masking important speech features. The same noise sample will not necessarily affect the intelligibility of all speakers equally. For instance, speakers with a peakier long term spectrum may, on average, be more resilient to energetic masking. Clearly, the greater the extent of energetic masking, the bigger the reduction in intelligibility. More subtly, the same degree of masking – measured in terms of spectro-temporal proportion masked – may have greater or lesser impact on intelligibility depending on *where* the masking occurs in relation to key features of the speech signal. These effects will be speaker dependent – heavy masking in a given frequency region may be more damaging to one speaker than another. For example, vocal tract length differences alter average formant frequencies, so a masker with energy at a frequency that occludes the second formant of one speaker, may, on average, fall harmlessly in the gap between two formants for another.

Unlike energetic masking, the effects of *informational masking* result from target and masker competition in more central portions of the auditory system (Durlach et al., 2003). For example, if there is a competing speaker present, listeners may experience difficulty focusing attention on the target speaker. In a series of simultaneous speaker intelligibility studies (Brungart, 2001; Brungart et al., 2001), Brungart and colleagues demonstrated that speech is more effectively masked by speech of the same gender, and that using the same speaker for both the target and masker produces an even greater masking effect. When the target and masker have the same level, the effect of informational masking is greatest. Here, we focus on energetic masking by employing stationary noise maskers that are not readily confusable with speech.

The current study explores the use of statistical modelling techniques adopted from automatic speech recognition (ASR) in the estimation of speaker intelligibility. Considering intelligibility from an ASR perspective, it is commonly observed that some speakers can use ASR systems with more ease than others. ASR systems are typically trained using data from a large number of speakers with the hope that the acoustic models produced will generalise to the unknown speaker attempting to use the system. Typically, adaptation techniques are then used to minimise error due to mismatch between the user and the learnt acoustic models (Woodland, 2001). However, ASR systems employing speaker adaptation are not good models for speaker intelligibility – for example, they can fail badly on non-native speech that humans may find highly intelligible (van Compernolle, 2001). In the current work, we use large amounts of speech from individual speakers to train speaker-specific models to avoid speaker adaptation issues. ASR systems also differ from humans in the way they respond to additive noise. Even when using speaker-specific acoustic models, it cannot be expected that ASR recognition results for a set of speakers in high levels of noise will correlate with human judgements of intelligibility. To produce more human-like results, ASR systems need to be adapted to base

their decisions on the portions of the speech signal that are not energetically masked.

In this paper we attempt to account for the variability in speaker intelligibility using a glimpsing model of speech perception (Cooke, 2006) which is built on statistical modelling techniques employed in ASR. This model takes the view that listeners process noisy speech by taking advantage of "glimpses" – spectro-temporal regions – in which the target is least affected by the background. The model uses missing data speech recognition techniques (Cooke et al., 2001) to predict the response of listeners to the noisy speech on a token by token basis. Estimates of intelligibility are made from the output of the adapted ASR system averaged over a large number of utterances. We characterise the model as being 'microscopic' as the underlying ASR system predicts responses to individual tokens – a feature that it shares with earlier models such as those of Ghitza (1993), Ainsworth and Meyer (1994) and Holube and Kollmeier (1996). Microscopic models contrast with 'macroscopic' models of intelligibility – such as the Articulation Index (Fletcher and Galt, 1950), the Speech Transmission Index (Steeneken and Houtgast, 1980) and the speech recognition sensitivity model (Musch and Buus, 2001) – which predict intelligibility without employing detailed acoustic models of speech, and without predicting listeners' detailed response to each utterance. Macroscopic models are based on long term statistics of the speech and noise, and they work to the extent that errors caused by smoothing out the detail do not cause bias when averaged over the material in the intelligibility test. Microscopic models have the potential to make more accurate estimates of intelligibility, but their evaluation requires a corpus that is both carefully controlled (and thus suitable for measuring intelligibility), and that has large amounts of data for each speaker (making it suitable for training statistical speech models). The current study has been made possible by the recent release of the Grid corpus (Cooke et al., submitted) which is further described in Section 2.2.

As well as providing a model of relative speaker intelligibility, the current study may be viewed as a test of the glimpsing model of speech perception. If the glimpsing model is a poor fit of the data then the intelligibility judgements are unlikely to be accurate. In Cooke (2006), the model was tested using a corpus of isolated vowel-consonant-vowel tokens collected by Shannon et al. (1999). A test set using 160 items from just 5 male speakers was used. In contrast, the Grid corpus employed in the current study is composed of connected 6-word utterances constructed from a vocabulary of 47 words, spoken by 34 different speakers of both genders. Crucially, there is sufficient training data to build *speaker-specific* acoustic models. The more natural style of the speech material, and increased perplexity of the recognition task provides a more stringent test of the glimpsing account.

The remainder of the paper is structured as follows. Section 2 describes the

listening tests that were conducted to estimate the intelligibility of the 34 Grid speakers in stationary noise conditions over a range of signal to noise ratios. Section 3 describes the acoustic analysis of the Grid speakers. Measurements are made of the degree to which simple acoustic parameters can predict speaker intelligibility. Section 4 applies the glimpsing model to estimate the degree of masking for each speaker at each noise level and examines the degree to which relative intelligibility can be explained in terms of masking in isolation from acoustic models of individual speakers. Section 5 presents the full ASR-based model and looks in detail at the extent to which it agrees with human intelligibility judgements.

2 Listening tests

2.1 Participants

Twenty native speakers of British English participated in the study. Listeners were students and staff at the University of Sheffield whose age ranged from 20 to 43 years (mean: 26.1 years, s.d. = 6.9). Students were paid for their participation. All listeners were screened for hearing loss (better than 20 dB hearing level in the range 250 – 8000 Hz). Ethics permission was obtained following the University of Sheffield Ethics Procedure.

2.2 Speech and noise materials

Sentences were drawn from the Grid corpus (Cooke et al., submitted) which provides common speech material for studies in speech perception and automatic speech recognition. Thirty four speakers (18 males and 16 females) provided 1,000 utterances each, producing a total of 34,000 sentences suitable for both intelligibility testing and for training automatic speech recognisers. Sentences in the Grid corpus are simple 6 word utterances such as "put red at X 4 now" and "set green with J 9 again". Sentences have a fixed syntax. Items in positions 2 (colour), 4 (letter) and 5 (digit) act as keywords to be identified. In Grid, the 4 colour choices are "red", "green", "blue" and "white", while 25 letters from the English alphabet are available ("W" is excluded due to its multisyllabicity) as are the 10 digits from "zero" to "nine" (see Table 1).

Speech-shaped noise whose spectrum matched the long-term spectrum of the entire Grid corpus was added to utterances at 11 SNRs: 6, 4, 2, 0, -2, -4, -6, -8, -10, -12 and -14 dB. These SNRs were chosen based on the results of a pilot experiment to cover the full intelligibility range. All utterances had initial and

Table 1
Structure of sentences in Grid corpus.

VERB	COLOUR	PREP.	LETTER	DIGIT	ADVERB
bin	blue	at	a-z	1-9	again
lay	green	by	(no 'w')	and zero	now
place	red	on			please
set	white	with			soon

final silence removed prior to the addition of the noise. An additional condition consisted of sentences presented without noise. For each of the 12 conditions and for each of the 20 listeners, an independent block of 100 utterances was drawn at random, without replacement, from the Grid corpus. Consequently, a total of 20 x 12 x 100 = 24,000 Grid utterances were used in this study.

2.3 Procedure

Listeners were tested individually in an IAC single-walled acoustically-isolated booth. Stimulus presentation and response collection was under computer control. Noisy utterances were scaled to produce a presentation level of approximately 68 dB SPL and were presented diotically over Sennheiser HD250 headphones. Listeners were asked to identify the colour, letter and digit spoken and entered their results using a conventional computer keyboard in which 4 of the non-letter/digit keys were marked with coloured stickers. Those keys representing colours were activated immediately following the onset of each utterance. As soon as a colour key was pressed, the 25 relevant letter keys were enabled, followed by the 10 digit keys. This approach allowed for rapid and accurate data entry: most listeners were able to identify a block of 100 utterances in 5–7 minutes. Listeners were familiarised with the stimuli and the task by identifying an independent practice set of 100 sentences prior to the main set. The order of presentation of the non-practice conditions (including the quiet condition) was randomised. Listeners identified the 13 blocks of 100 utterances over two sessions of around 40 minutes each on separate days.

2.4 Results

Listeners' responses were scored in terms of colours, letters and digits correct as well as all keywords correct. Figure 1 shows identification rates for these 4 measures as a function of SNR. As expected, the number of choices for each keyword influences the results, with better identification rates at all SNRs for colours, digits and letters respectively.

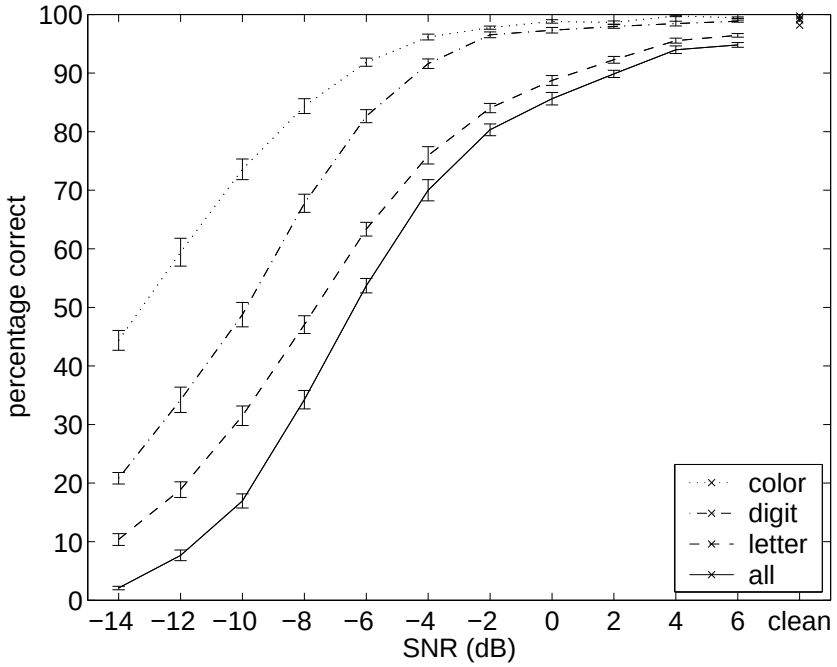

Fig. 1. *Percentage of colour, digit and letter keywords recognised correctly, averaged across listeners as a function of SNR. The solid line shows the percentage of utterances in which all the keywords were recognised correctly. Error bars here and elsewhere denote +/- 1 standard error.*

For the purpose of this study, overall keyword intelligibility was measured as the percentage of keywords correct (i.e. the average of the colour, letter and digit correct scores). Figure 2 plots overall intelligibility separately for male and female speakers. A repeated-measures ANOVA with one within-subjects factor (noise level) and one across-subjects factor (gender) demonstrated a small but significant effect of gender ($F(1,32)=5.00$, $p < 0.05$, $\eta^2 = 0.14$): female speakers were more intelligible than males at most SNRs, a difference equivalent to about 1 dB of noise.

For some of the subsequent analyses, the 12 conditions have been grouped into 3 sets: a low noise condition (clean and 6, 4, 2 dB SNR), a medium noise condition (0, -2, -4 and -6 dB SNR) and a high noise condition (-8, -10, -12 and -14 dB SNR). Identification rates for each of the keyword options for the three noise level ranges are displayed in Figure 3. There is a wide spread in the identification rate of keywords, particularly across the set of letter tokens. For example, in the high noise condition, the identification rate for letters 'b' and 'v' is little over chance level (4%), whereas the letter 'r' is recognised correctly nearly 50% of the time. A more detailed investigation of confusions patterns revealed that most occur between members of the 'e'-set ('b','c','d','e','g','p','t','v'), with confusions between 'b' and 'v' being particularly common. In addition, 'm' and 'n' are frequently confused.

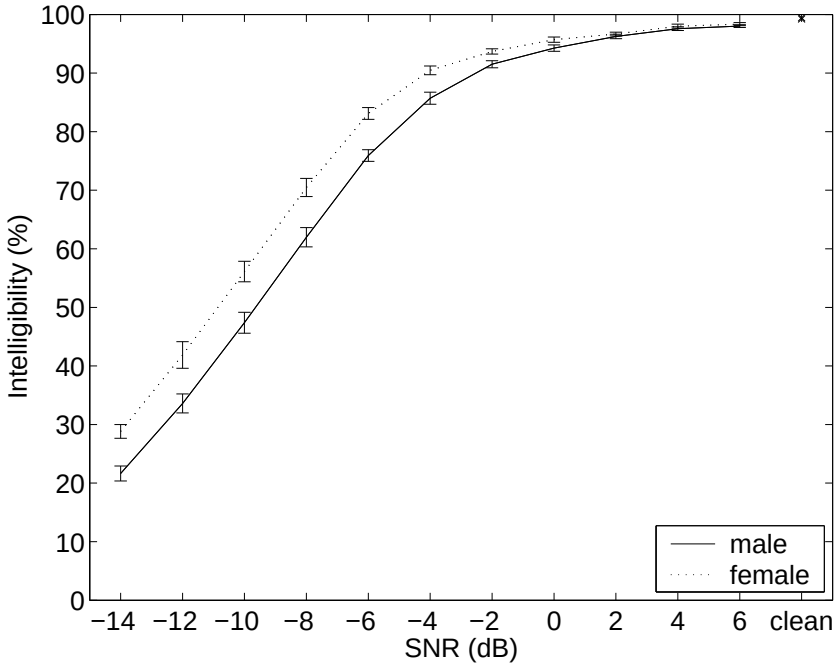

Fig. 2. *Overall keyword intelligibility of male (solid) and female (dashed) speakers as a function of global SNR.*

Figure 4 shows overall keyword intelligibility of individual speakers in the three grouped noise levels. The spread of intelligibility across speakers is large, with values in the high noise case ranging from as high as 68% (speaker 18) to as low as 24% (speaker 1). Mean intelligibilities are 98.0%, 88.7% and 45.3% for the low, medium and high level noise bands respectively, with corresponding standard deviations of 1.2%, 6.5% and 10.5%. The spread in the cleaner conditions is reduced because many utterances are recognised without error, so there is a ceiling effect that becomes more significant as the SNR increases.

3 Acoustic analysis

3.1 Acoustic measurements

To investigate possible factors underlying the intelligibility of Grid sentences in quiet and in noise, a series of acoustic measures was estimated for individual sentences and for each speaker. For each utterance, mean fundamental frequency (F0) and overall duration were computed. F0 estimates and binary voicing decisions were provided at 10 ms intervals using an autocorrelation-based method (Boersma, 1993) implemented in the Praat program (Boersma and Weenink, 2005). Durations were derived from utterance endpoints computed via forced-alignment of the original Grid utterances. Since Grid sen-

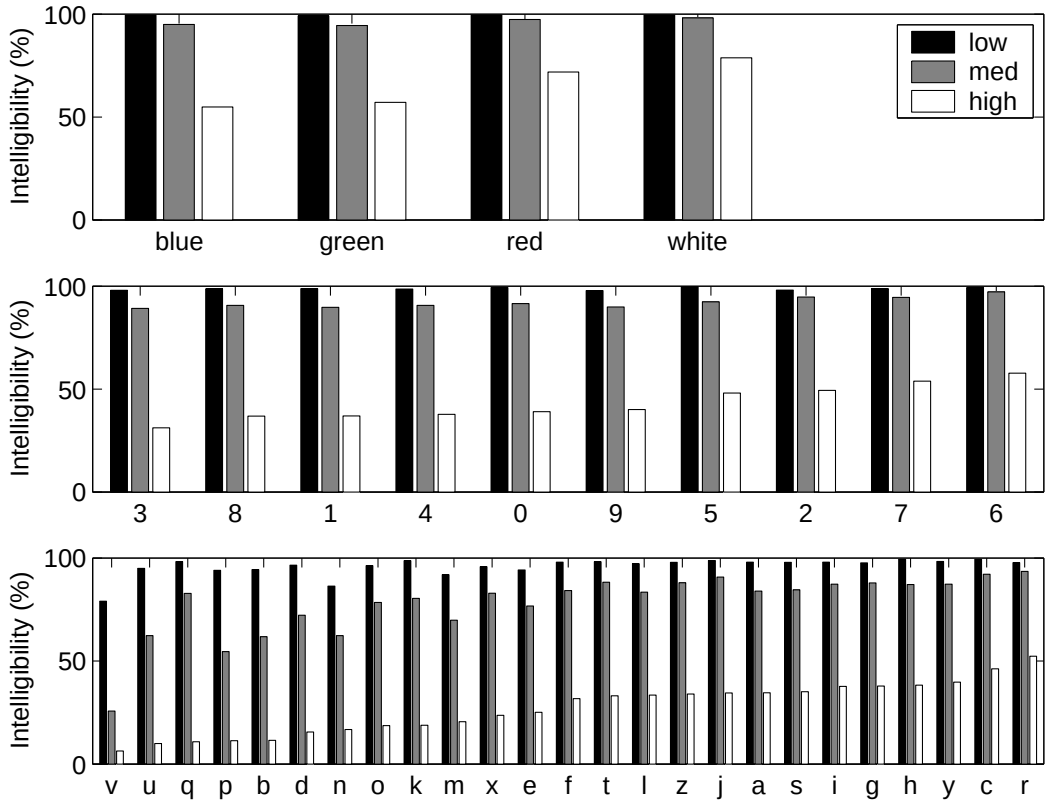

Fig. 3. *Intelligibility of the individual colour (top), digit (centre) and letter (bottom) tokens in either the low (black), medium (grey) or high (white) noise condition. Tokens are arranged by order of intelligibility in the high noise condition.*

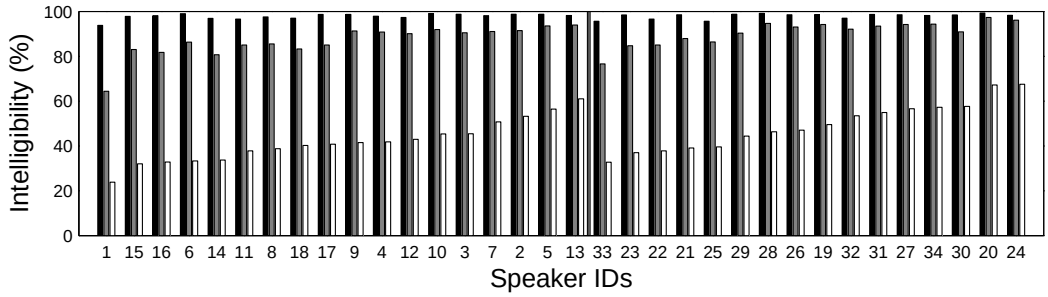

Fig. 4. *Intelligibility of each of the 34 speakers in the Grid corpus in either the low (black), medium (grey) or high (white) noise condition. Male speakers are shown on the left while female speakers are on the right. Within each gender, speakers are arranged by order of intelligibility in the high noise condition.*

tences are of the same length, duration can also be interpreted as speech rate.

In addition to the F0 and duration measures, a vocal tract length (VTL) warping factor was estimated. Using phone-level alignments, all instances of the high vowels /i/, /ɪ/ and /ɛ/ in the Grid corpus were identified for each speaker independently. These sounds were chosen because many examples exist (around 3,000 per speaker), allowing robust estimation of VTL warping

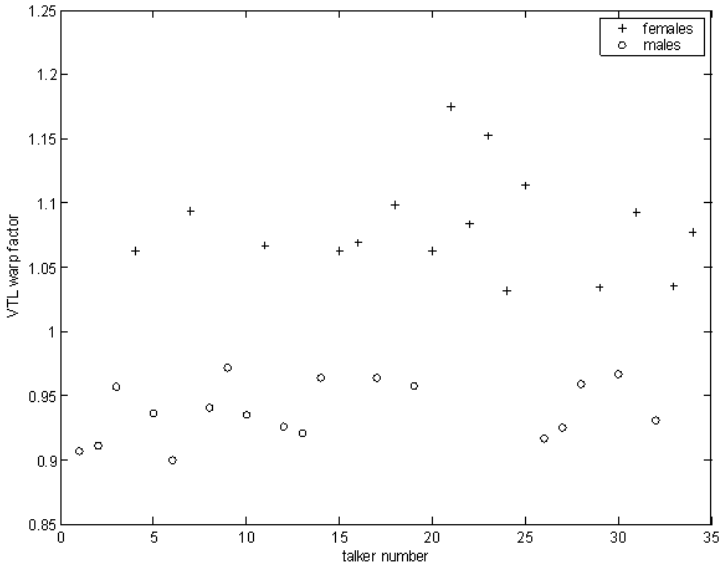

Fig. 5. *VTL warp factors for each speaker in the Grid corpus, estimated from the vowels /i/, /ɪ/ and /ɛ/. (o) males, (+) females.*

factors. Frequencies for the first 3 vowel formants ($F1$, $F2$ and $F3$) were estimated at the mid-point of the time interval corresponding to each vowel instance. The Burg algorithm (Burg, 1975) implemented in Praat was used for formant analysis. For each speaker and each of the 3 vowels independently, median values for $F1$, $F2$ and $F3$ across all instances were computed. To provide a reference point for formant frequency warping, a median for each formant and each of the 3 vowels across all speakers was calculated. A multiplicative factor, α_v, which, when applied to each formant, minimises the distance between the formant frequency estimates ($F1_s$, $F2_s$, $F3_s$) of an individual speaker s and those of the 'average' speaker ($F1_a$, $F2_a$, $F3_a$) was computed for each vowel v independently using Equation 1,

$$\alpha_v = \exp((\log(F1_s/F1_a) + log(F2_s/F2_a) + log(F3_s/F3_a))/3) \quad (1)$$

Similar VTL warp factor estimates were produced for each vowel using this procedure. The median of the 3 estimates ($\alpha_{/i/}, \alpha_{/I/}, \alpha_{/E/}$) was taken as the final VTL warp factor estimate. Warp factors for the 34 speakers are shown in Figure 5. A clear separation between the male and female speakers can be seen. Note that the 'average' speaker has a warp factor of unity. In the subsequent discussion, the abbreviation VTLW is used to refer to warp factor estimates, but it should be noted that larger values of VTLW correspond to smaller VTLs.

A pairwise correlation analysis for the 3 measures introduced above, conducted over all 34,000 sentences in the Grid corpus, showed a strong positive

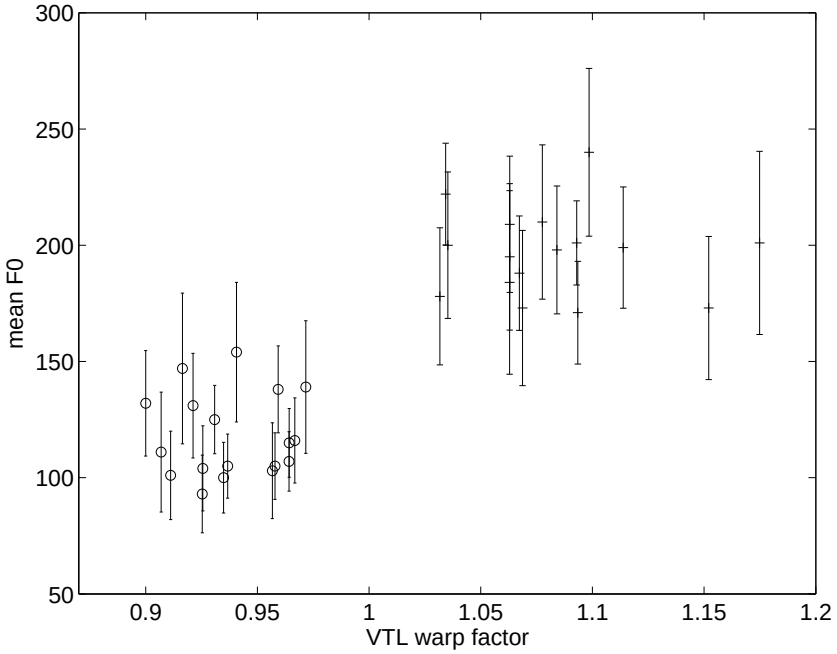

Fig. 6. *VTL warp factors plotted against mean utterance F0 for each speaker in the Grid corpus. (o) males, (+) females.*

relationship between mean utterance F0 and VTLW for the speaker producing that utterance (correlation coefficient $\rho(VTLW, F0) = 0.82$). However, this correlation was almost entirely due to speaker gender, as shown in Figure 6. Perhaps surprisingly, when utterances from male and female speakers were analysed separately, no significant correlation was found between the two parameters (($\rho_{males}(VTLW, F0) = -0.03$, $\rho_{females}(VTLW, F0) = -0.05$). Hence, for both the male and female groups, an increase in VTLW does not result in an increase in mean F0 per utterance. Correlations between duration and the other two measures were also found to be non-significant ($\rho_{males}(VTLW, duration) = 0.07$, $\rho_{females}(VTLW, duration) = -0.05$, $\rho_{males}(F0, duration) = 0.17$, $\rho_{females}(F0, duration) = -0.01$).

3.2 Influence of VTLW, mean F0 and duration on intelligibility

The effect of VTLW, mean F0 and durational differences on intelligibility was examined for the male and female speakers separately. For each of the measures, the set of sentences heard was partitioned into three equal-sized groups based on the measure value. To illustrate this process in the case of the mean F0 measure, two breakpoints were chosen such that all sentences with a mean F0 smaller than the first breakpoint formed a single subset, while all those with a mean F0 falling between the two breakpoints constituted the second subset, with the remainder forming the third subset. This process of aggregation was used to allow robust estimation of intelligibility scores per

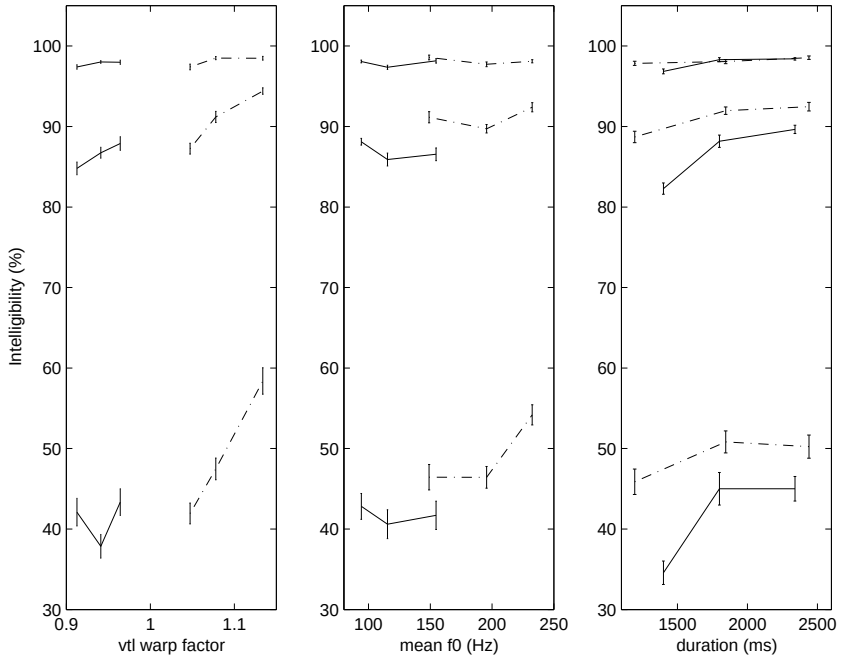

Fig. 7. *The influence of vocal tract length, mean F0 and duration on intelligibility for male (solid lines) and female (dotted lines) speakers. In each panel, one pair of lines for each of the 3 noise level groupings is shown.*

listener. Note that this process produced different groupings of utterances for each of the 3 measures. As before, noise conditions were grouped into 3 levels (low, medium and high noise) for ease of interpretation.

Figure 7 shows listeners' mean intelligibility scores as a function of differences in VTLW, mean F0 and duration for the male and female speakers, at the 3 grouped noise levels. For each measure and gender, a two factor repeated-measures ANOVA was carried out with measure subset and noise level as within-subject factors. In all cases, the effect of noise level was highly significant ($p < 0.001$), but of most interest was the effect of differences in the measured parameter.

For male speakers, a significant interaction between noise level and VTLW was found ($F(4,16)=9.42$, $p < 0.001$, $\eta^2=0.702$). The effect of VTLW was significant for medium ($F(2,18)=4.15$, $p < 0.05$, $\eta^2=0.315$) and high ($F(2,18)=12.11$, $p < 0.001$, $\eta^2=0.574$) noise levels. However, the pattern at the high noise levels was non-monotonic with no significant increase with increasing VTLW. At medium noise levels, pairwise comparisons indicated that the increase between the lowest and highest terciles was significant ($p < 0.05$). For the females, the effect of VTLW differences was marginally significant at the lowest noise level ($F(2,18) = 4.7$, $p < 0.05$, $\eta^2 = 0.342$) but highly significant at the medium ($F(2,18) = 68$, $p < 0.001$, $\eta^2 = 0.882$) and high ($F(2,18) = 62$, $p < 0.001$, $\eta^2 = 0.872$) noise levels. Pairwise comparisons of VTLWs in the individual noise conditions indicated that while at the low noise level the lower two VTLW

values were marginally different (p=0.026), at the medium and high noise levels, all VTLW values differed ($p < 0.001$). Given the relatively small number of speakers, it would be unwise to claim a clear effect of increasing VTLW on intelligibility for the male speakers. However, for the female speakers, the results do indeed suggest that vocal tract length plays an important role in intelligibility, especially in high noise conditions.

A very similar picture emerged for the mean F0 measure, with a main effect of mean F0 for the males just failing to make significance ($p = 0.07$), although pairwise analyses in the three noise levels independently showed significant effects of F0 ($p < 0.05$) in the low and medium noise conditions, with low F0s advantaged. However, the female speakers showed a highly significant effect of mean F0 ($F(2,18) = 23$, $p < 0.001$, $\eta^2 = 0.720$) and interaction with noise level ($F(4,16) = 16$, $p < 0.001$, $\eta^2 = 0.805$). No effect of mean F0 was found at the low noise level ($p = 0.175$) but a significant effect was found at both medium and high noise levels due to the difference between the higher pair of mean F0 values ($p < 0.001$). Consequently, it appears that sentences with the highest mean F0 values (upper tercile of the female speakers, mean F0 > 200 Hz) are most intelligible and there is some marginal evidence that sentences with the lowest mean F0 values (male speakers, mean F0 < 100) are also more intelligible than those with higher F0s produced by males.

Clear effects of durational differences can be seen. For both male and female speakers, there is a highly significant main effect of duration (males: $F(2,18) = 179$, $p < 0.001$, $\eta^2 = 0.952$; females: $F(2,18) = 27$, $p < 0.001$, $\eta^2 = 0.747$), which is also significant at all noise levels individually, though only marginally so for the female speakers at the lowest noise level, presumably due to ceiling effects. For the medium and low noise levels, the difference is entirely due to the durational increase from the lower to the middle tercile. This suggests that utterances with a high speech rate create greater difficulties for keyword identification in noise, but that the moderate rates present no more difficulty than the slowest rates.

Although utterance intelligibility appears to be related to F0 and duration, this does not imply that these acoustic parameters are useful for predicting relative *speaker* intelligibility. To test whether the effect of VTLW, F0 and duration observed across *utterances* produces an effect of intelligibility across *speakers*, each of the three acoustic parameters measured on a per-speaker basis (i.e. mean F0 per speaker, mean duration per speaker and speaker VTLW) was tested for correlation with speaker intelligibility. Correlations were measured separately at each of the 12 SNR levels. Correlations were computed across the full set of 34 speakers, and separately across either the 18 male speakers, or 16 female speakers. Results of this analysis are shown in the top row of Figure 11. Across all speakers, VTLW is loosely correlated with intelligibility across the SNR range -14 dB to -4 dB. However, within this SNR range

there is no significant intelligibility/VTLW correlation within male speakers, and only very marginal correlation within female speakers. The effect observed across all speakers is likely to arise from the fact that VTLW predicts gender and female speakers are more intelligible than male speakers at low SNR (possibly for reasons unrelated to vocal tract length). Within the set of male speakers, duration is correlated with intelligibility in the range -8 dB to 2 dB, but not at 0 dB. Duration does not appear correlated with intelligibility amongst the female speakers. This lack of effect among female speakers occurs in spite of the fact that the range of durations in the female set is actually slightly greater than that in the male set (Figure 7). Fast talking female speakers do not have the reduced intelligibility observed with fast talking male speakers. In summary, across the noise conditions studied, none of the acoustic parameters is strongly correlated with intelligibility. This implies that neither F0, VTL or duration can be used alone to make a useful prediction of relative speaker intelligibility. The fact that the strong effects observed in per utterance analysis are not reflected in the per speaker results, suggests that within-speaker differences in mean F0 and duration are large compared with between-speaker differences.

4 Glimpse analysis

4.1 Acoustic measurements

Speech is sparsely encoded with energy concentrated in compact regions of the spectro-temporal plane. Even at highly unfavourable SNRs, there will be local regions where the speech stands clear of the noise floor. The size, shape and spectro-temporal position of these glimpses will be dependent on the characteristics of the speaker. For example, a speaker with a peakier long term spectrum is likely to produce more glimpses than a speaker with a flatter spectrum. Availability of reliable speech glimpses is likely to be a contributing factor to the intelligibility of the speech (Cooke, 2003, 2006). To test this hypothesis, a model of glimpsing was applied to each speaker, and measurements were made of the percentage of the spectro-temporal information glimpsed, a quantity we call *visibility*.

Location of speech glimpses followed Cooke (2006). The model proceeds by constructing an auditory spectrogram representation of the speech data. This involves passing the input speech signal through a bank of 64 gammatone filters with centre frequencies ranging from 50 Hz to 8 kHz linearly spaced on an ERB-rate scale. Within each channel, the Hilbert envelope is computed and subjected to a leaky integrator with an 8 ms time constant. Finally, the output consists of 64 point 'auditory spectra' sampled at 100 Hz. This representation

was computed for both the clean Grid utterances and the speech-shaped noise used in the construction of noisy tokens. By comparing the spectra for the unmixed speech and noise signals, a local SNR estimate can be computed. Spectro-temporal points in the noisy utterance are marked as reliable if the local SNR at that point is greater than a threshold of T dB on a log energy scale. A glimpse is then defined to be a reliable region whose size is greater than N spectro-temporal points, where the size of the region is measured in terms of its 4-connectivity. We then define the *visibility*, V, of each utterance, to be,

$$V = 100 * \frac{Area_{glimpsed}}{Area_{spectrogram}}\% \tag{2}$$

where $Area_{glimpsed}$ is the count of the number of spectro-temporal points occurring within all the glimpses, and $Area_{spectrogram}$ is the 'area' of the auditory spectrogram, computed as the number of frequency channels multiplied by the number of frames. The glimpse location model has two free parameters, the SNR threshold, T, and the minimum glimpse size, N. These values are tuned empirically – as described in Section 5.2 – by minimising the distance between the mean intelligibility curves shown in Figure 1 and the corresponding estimates produced by the full ASR-based glimpsing model (presented in the next section). This tuning procedure determined suitable values for T and N to be 3 dB and 5 respectively.

4.2 Influence of visibility on intelligibility

The effect of visibility on intelligibility was studied for the male and female speakers separately. For both groups, sets of low, medium and high visibility utterances were created using the same procedure as that used to partition the data for the acoustic analyses: viz. all utterances from all speakers were ordered by their mean visibility, and then the utterances were partitioned into three equal-sized sets. The average intelligibility of each set was computed. This analysis was repeated for each global SNR.

Figure 8 shows average intelligibility versus average visibility for the low visibility and high visibility sets for female and male speakers at global SNRs of -14, -12, -10, -8, -6 and -4 dB. More favourable global SNRs were not used since performance is already near to ceiling. The first point to note is that high intelligibility can be achieved despite a relatively small amount of the spectrum being visible. At -4 dB less than 10% of the spectrum is visible but the intelligibility has reached 80 to 90%. Further, only 2% of the spectrum is required to recognise half the words correctly. This is evidence of the high degree of redundancy in the speech signal for this limited perplexity task. Clearly,

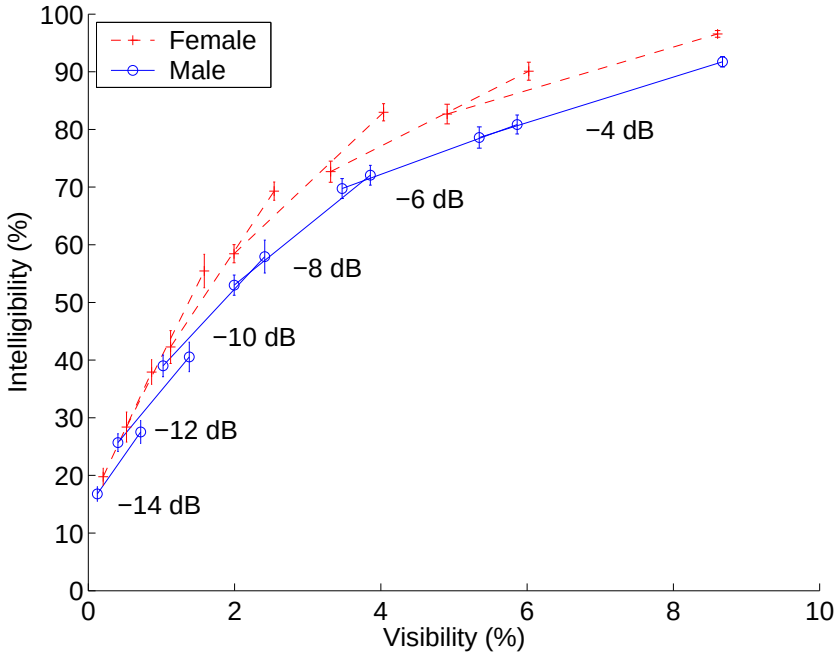

Fig. 8. *Intelligibility of low and high visibility utterances, for male (o) and female (+) speech, at the SNRs marked.*

as the SNR is increased the visibility of the target increases and there is an accompanying rise in intelligibility. More surprisingly, within a single SNR there is a large difference between the mean visibility of the low visibility and high visibility sets. So, while holding global SNR constant, differences in spectral characteristics of the utterances have an effect on visibility equivalent to around 2 dB of global SNR variation. The intelligibility of the high visibility utterances is significantly ($p < 0.001$) greater than that of the low visibility utterances for both male and female speakers at all SNR levels in the range -14 to -4 dB. At each SNR, the visibility of the male and female utterances is approximately equal. However, the female utterances have consistently higher intelligibility for a given degree of visibility. Also, the gain in intelligibility between the low and high visibility sets is greater for the female speakers than for the male, i.e. the same amount of extra information is adding more to the intelligiblity of the female speech.

Following the analysis of the previous section, the visibility data was also analysed on a per-speaker basis. The correlations between speaker visibility and speaker intelligibility in the noise range -14 dB to -4 dB are shown separately for the male and female speakers in Table 2. Significant correlations are marked. A plot of the correlations over the full SNR range – and across the combined male/female data – is shown in the 2nd row of Figure 11.

Considering female speakers first, it can be seen from Figure 11 that visibility is highly correlated with intelligibility, especially in the range -12 to -8 dB. At

Table 2
Correlation between speaker visibility and speaker intelligibility for male and female speakers at each of the 6 lowest SNRs. Significant values ($p<0.05$) are marked with * while highly significant values ($p<0.005$) are marked with **.

SNR	-14 dB	-12 dB	-10 dB	-8 dB	-6 dB	-4 dB
male	0.48 *	0.49 *	0.53 *	0.38	0.40	0.48 *
female	0.75 **	0.91 **	0.83 **	0.81 **	0.75 **	0.66 **

the -14 dB point the correlation is reduced. At this extreme SNR there is a performance floor effect where human performance reaches chance levels for the least intelligible speakers. Correlation will therefore be reduced because there is no significant difference in the intelligibilities of a subset of speakers. The model performs less well at higher SNRs. At high SNRs where all speakers have high visibility, differences in intelligibility are not likely to be governed by the small inter-speaker differences in absolute visibility. As the speech approaches the clean condition, intelligibility is probably more closely related to the intrinsic intelligibility – or clarity – of the speaker. As reviewed in the introduction, intrinsic intelligibility appears to be related to factors such as the consistency of pronunciation, and the size of the vowel space, variables that are perhaps quite independent of visibility. Amongst male speakers, correlations between visibility and intelligibility are much weaker. Significant correlations occur only at -14, -12, -10, -4 and 2 dB and these correlations only just reach $p = 0.05$. This result is surprising considering the strength of the effect that visibility has on intelligibility, as seen in the per-utterance based analysis (Figure 8). One explanation would be that although there is a similar spread of visibilities across utterances of male and female speech, there is less *inter-speaker* variability of visibility for male speech than for female speech. However, the data does not support this explanation, because at the lowest SNRs – where the female visibility/intelligibility correlation are high – there is a larger spread of visibilities across male speakers than across female speakers.

5 Statistical modelling

In the previous section it was seen that at low SNRs there is a relation between visibility and intelligibility for female speech but not for male speech. At high SNRs, visibility is not a good predictor of intelligibility for either gender. Furthermore, female speech is generally more intelligible than male speech of equal visibility. Clearly, visibility alone is not sufficient to model speaker intelligibility. This is not surprising as studies using clean speech (i.e. 100% visible) have shown that speakers have a range of intrinsic intelligibility (Bradlow et al., 1996; Hazan and Markham, 2004). It is not sufficient simply to calculate how much of the signal is glimpsed. The relative usefulness of glimpses must also

be considered. One way to achieve this is to relate the glimpses to prior knowledge of speech in the form of acoustic models. In this section, we implement the full glimpsing model described in Cooke (2006) which attempts to predict average listener response to noisy speech using statistical ASR techniques to interpret the glimpses located by the energetic masking model in the auditory front-end.

5.1 The ASR-based glimpsing model

Statistical models were constructed to represent each of the 34 speakers in the corpus. The models were based on the 64-channel gammatone filterbank representation described in the previous section. The 64-dimensional spectral vectors were supplemented with their 1st order temporal derivatives – computed using linear regression over a 5 frame window – to produce a 128-dimensional feature vector.

For each speaker in the corpus a set of word-level hidden Markov models (HMMs) was trained from the auditory spectrogram data. Each word model was represented using two states per phoneme, i.e. the number of states varied between two (the letter 'E') and ten ('seven'). Each state had two transitions; a self transition and a transition to the adjacent state. A three state silence model was used to represent the silence period before and after the utterance, and a single state model was used to model optional short pauses between words. The short pause model had a transition between its non-emitting start and end states allowing it to be skipped when it was not required. Each state was modelled using a Gaussian mixture model with 5 diagonal covariance components. The following 'round-robin' training procedure was used to maximise the amount of training and test data. The 1,000 utterances were randomly divided into five test sets containing 200 utterances each. For each test set a corresponding set of models was built using the remaining 800 utterances as training data. Training proceeded from a 'flat start' where all model states were initialised with a single Gaussian whose mean and variance were computed across the complete data set. Mixture splitting was employed to increase the number of mixture components to two, then three, and finally five.

The statistical models were employed as described in Cooke (2006), using the energetic masking model described in Section 4 to estimate which spectro-temporal regions may be regarded as reliable. This information was represented as a binary spectral-temporal mask in which 1's indicate spectro-temporal regions in which the speech signal is glimpsed, and 0's indicate spectro-temporal elements that are effectively masked. Given this mask, the noisy utterance can be recognised using the speaker dependent model of clean speech using the bounded marginalisation missing data technique described

in (Cooke et al., 2001).

5.2 Tuning model parameters

The glimpse model has two free parameters: T, the local SNR threshold for glimpse detection, and N the minimum glimpse area. These parameters were tuned using a subset of 4 speakers (two male, and two female) drawn randomly from the 34 speakers in the grid corpus. A grid of parameter settings was tested with values of T being -7, -5, -3, 0, 3, 5, 7 or 9 dB and with N being either 1, 5 or 25. For each parameter setting, the recognition results produced by the model were compared with those of listeners on the same subset of speakers. The distance between listener and model performance was measured by computing the mean absolute difference between the points on the global SNR versus recognition correctness curves for the colour, digit and letter tokens. The best fit was found at $T = 3$ dB and $N = 5$. Different subsets of 2 male and 2 female speaker were tested, and all produced the same best-fit N and T values. At this granularity and for this task, the tuning of N and T appears to be speaker-independent.

Figure 9 compares model and listener performance across all 34 speakers. At the best-fit tuning, performance curves for the model have a similar shape to those of listeners. However, the model underestimates listener letter recognition performance at high SNRs. Closer inspection shows that the model makes more confusions amongst acoustically confusable letter pairs – particularly, 'm/n' and 'b/v'. However, here we are principally interested in *relative* speaker intelligibility using the parameter values which best fit the overall performance.

As the local SNR threshold, T, is decreased below the 3 dB point, the size of the glimpses becomes larger but the glimpses become contaminated with more noise. In this region, model fits become poorer as it starts to perform *better* than humans at very low global SNRs. For example, letter recognition at the -14 dB global SNR rises as the local SNR threshold is reduced and peaks at about 40% correct at a local SNR threshold of -5 dB. At the same global SNR human letter recognition is only 10%. That humans cannot achieve better performance at low global SNR suggests that it may not be possible to detect glimpses where the local SNR is much below 0 dB, i.e. the model can simulate values of T below 0 dB using *a priori* information, but these values may not be attainable in practice. An alternative explanation for the best-fit local SNR threshold not being lower is motivated by the observation that at higher global SNRs, where larger glimpses can be found at a given local SNR threshold, the trend is reversed – reducing the local SNR threshold actually *reduces* recognition performance. It appears that when the global SNR is high,

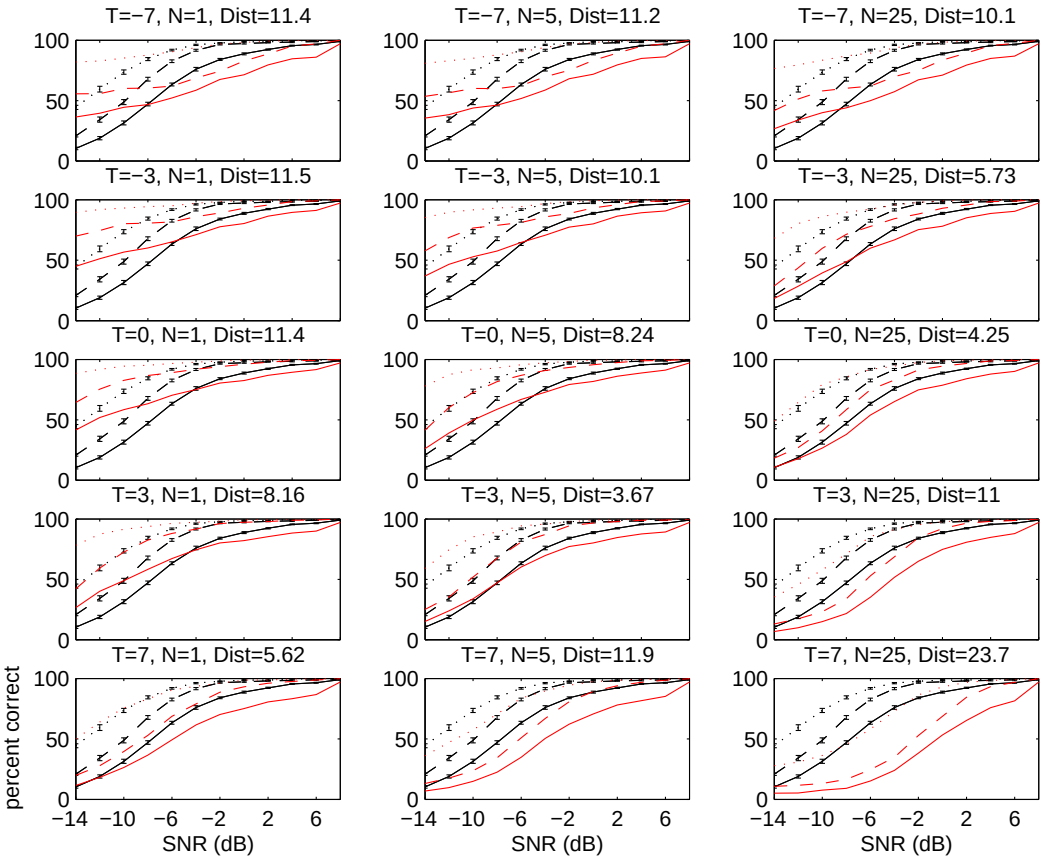

Fig. 9. *Comparison of listener and model recognition scores for colour (dotted), digit (dashed) and letter (solid) tokens across all SNR for a range of model parameters. Scores for listeners are those shown with +/- 1 standard error bars. Parameter settings and distance score are shown in the title of each subplot. The model best matches listener recognition data when $T = 3$ dB and $N = 5$.*

it is better to have a high value of T so that the glimpses are kept noise-free, while for high-noise conditions, it is preferable to operate at a low value of T so that the glimpses are larger, but at the expenses of the glimpses being less reliable. The T value of 3 dB is perhaps a reasonable compromise that works well across the range of global SNRs at which speech recognition can be usefully employed.

5.3 Results

The model was applied to produce recognition results for each speaker at each SNR. Here, intelligibility and ASR-based scores were computed as the average proportion of keywords correctly identified, expressed as a percentage and then arcsine-transformed. As before, recognition scores for each speaker were averaged across the low, medium and high noise level groupings. Correlations

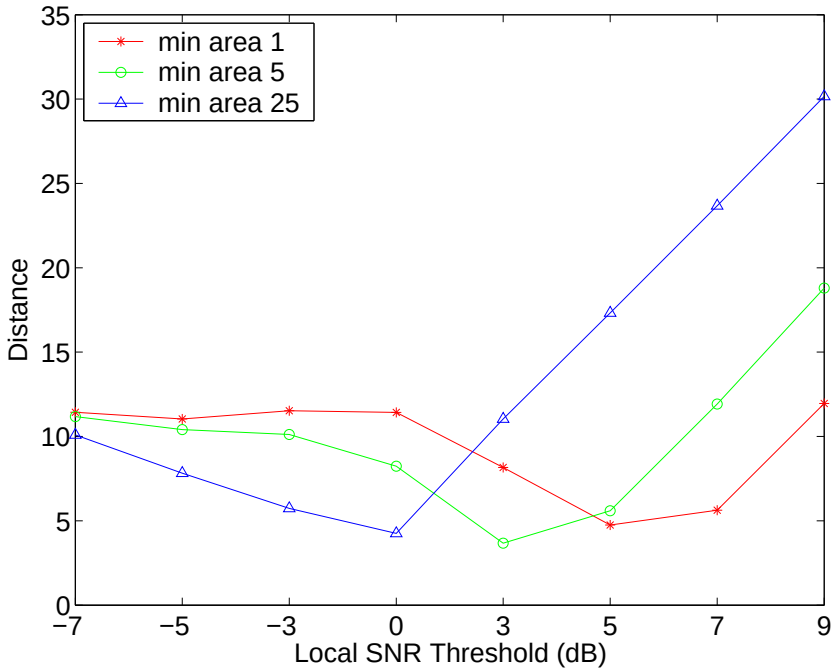

Fig. 10. *Effect of local SNR threshold, T, and minimum glimpse area, N, on the distance between ASR and listener recognition performance curves. The best fit to the listener data occurs when T is 3 dB and N is 5.*

between the transformed ASR-based recognition scores and the transformed human recognition scores across all 34 speakers for each noise level band can be seen in Figure 12. ASR-based and HSR scores were normalised to have equal mean and variance before plotting – a transformation which does not affect the correlation but which makes the figure easier to interpret. The correlation coefficients for the low, medium and high noise conditions are 0.60, 0.77 and 0.93 respectively. The model produces the best fit to the human data in the highest noise conditions. In the medium noise condition the correlation is high, but there are a subset of speakers for which the fit is poor – most notably the male speakers 1, 5, 8 and 17.

As per the analysis of the acoustic measures and visibility, correlations were computed at individual global SNRs both separately across either male or female utterances, and jointly across all speakers in the corpus. Results are shown in the second row of Figure 11. Correlation between visibility and intelligibility is plotted on the same axis for comparison. It can be seen that the full ASR-based model consistently produces results that have much greater correlation with listener performance than does the simple visibility measure. While visibility alone is only a strong predictor of the pattern of intelligibility for female speaker at low SNRs, the full glimpsing model can also predict the relative intelligibility of male speakers at low SNRs, and of female speakers at high SNRs. Reduced correlations at extreme SNR values are due to floor and ceiling effects: In the clean condition the high quality of the speech recordings,

and the small vocabulary nature of the recognition task, meant that listeners made very few errors, so differences between speakers are not significant. At the extreme -14 dB SNR point many of the speakers become totally unintelligible, so again there is no significant difference between a subset of the intelligibility scores.

Although the model performs well across a range of SNRs, the correlations are noticeably poorer in the region -2 to 0 dB. It is possible that this is due to effects of informational masking which are not accommodated in the current model. Although stationary noise is normally considered to have no informational masking potential and is a poor match to the spectrum of any particular speech sound, there may exist small spectro-temporal glimpses of noise that are well matched to some part of the speech models. These confusions will occur most often at 0 dB SNR where the speech shaped noise masker has a similar level as the target speech. The degree of confusion between the speech and the noise may be speaker dependent. For example, the effect is likely to be larger for speakers with a long term spectrum close to the average spectrum used for producing the speech shaped noise. The glimpsing model does not account for these effects. Glimpses of speech are located using *a priori* information, so regions of the background are never erroneously labelled as belonging to the speech source.

Finally, it is interesting to note that since the ASR-based model is 'microscopic', it can also be used to predict the relative intelligibility of individual tokens averaged across speakers. Figure 13 compares the mean and variance normalised arcsine-transformed ASR-based and HSR results for letters in each of the three noise level bands. The bar on the left of each pair represents the HSR result, and the bar on the right the ASR-based result. Correlation again increases as the noise level is increased, taking values of 0.78, 0.83 and 0.85 for low, medium and high noise levels respectively. These correlations are all highly significant. In the noisiest condition the general pattern of intelligibility is closely matched, with only a few letters not fitting. Most noticeably, the intelligibility of 'a' is underestimated, and that of 'o' is overestimated.

Correlations are not as high for the digit and colour keywords, possibly because the notion of 'guessing' is different in the model and listeners. When listeners are forced to guess, some have a tendency to repeatedly select the same response. For different reasons, when the model is faced with insufficient evidence, it may consistently select the same token, based on such criteria such as ease-of-masking. For instance, there is some evidence in the letter responses that listeners pressed 'A' more frequently than expected while the model output 'O' overly often. These artefacts will have little effect on correlation scores when there are a large number of response categories, but make a bigger difference when there are fewer tokens, as is the case for the digits and colours.

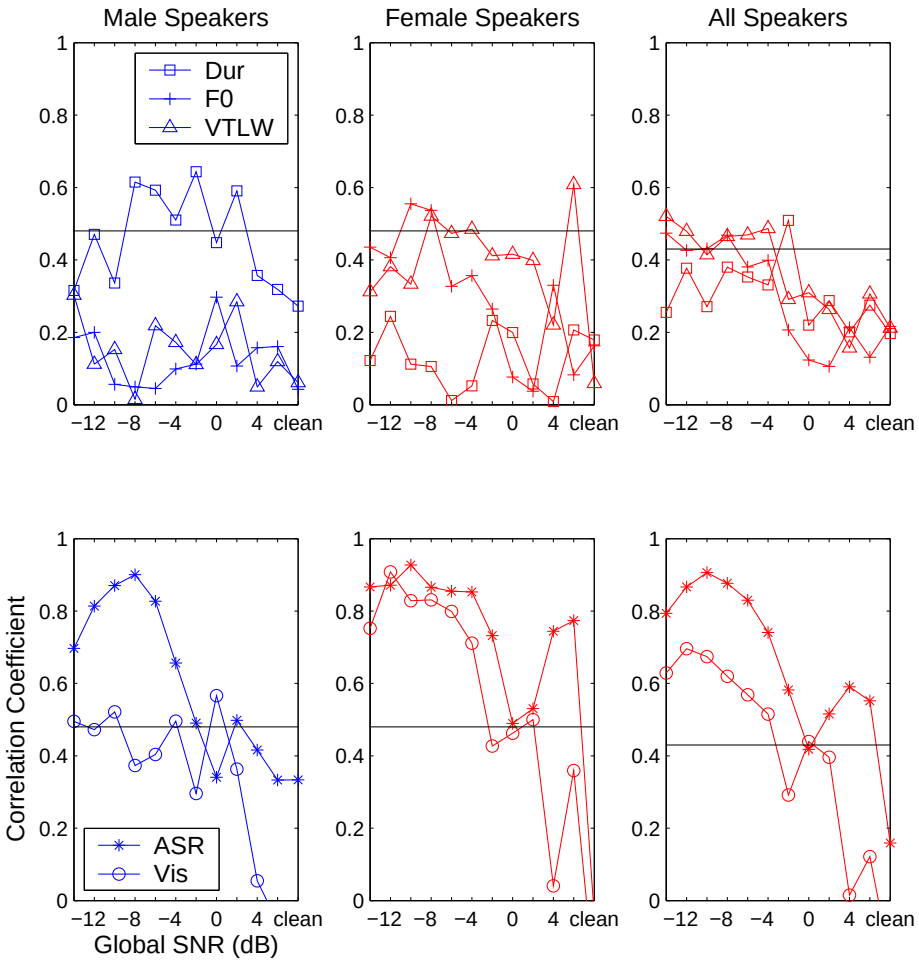

Fig. 11. *Summary of correlations.* ***Top row:*** *Correlations between intelligibility and acoustic factors – duration (squares), F0 (crosses) or VTL warp factor (triangles) – measured across male speakers (left), female speakers (centre) or all speakers (right) at each global SNR.* ***Bottom row:*** *Correlations between intelligibility and visibility (o) and between intelligibility and ASR-based scores (*) measured across male speakers (left), female speakers (centre) or all speakers (right) at each global SNR. Correlations above the horizontal bars are significant at the $p<0.05$ level.*

6 Discussion

6.1 Principal findings

This study compared behavioural performance on a multispeaker speech-in-noise task with that of a model inspired by automatic speech recognition techniques. Listeners identified 3 keywords in simple 6-word sentences in speech-shaped noise spoken by one of 18 male or 16 female speakers.

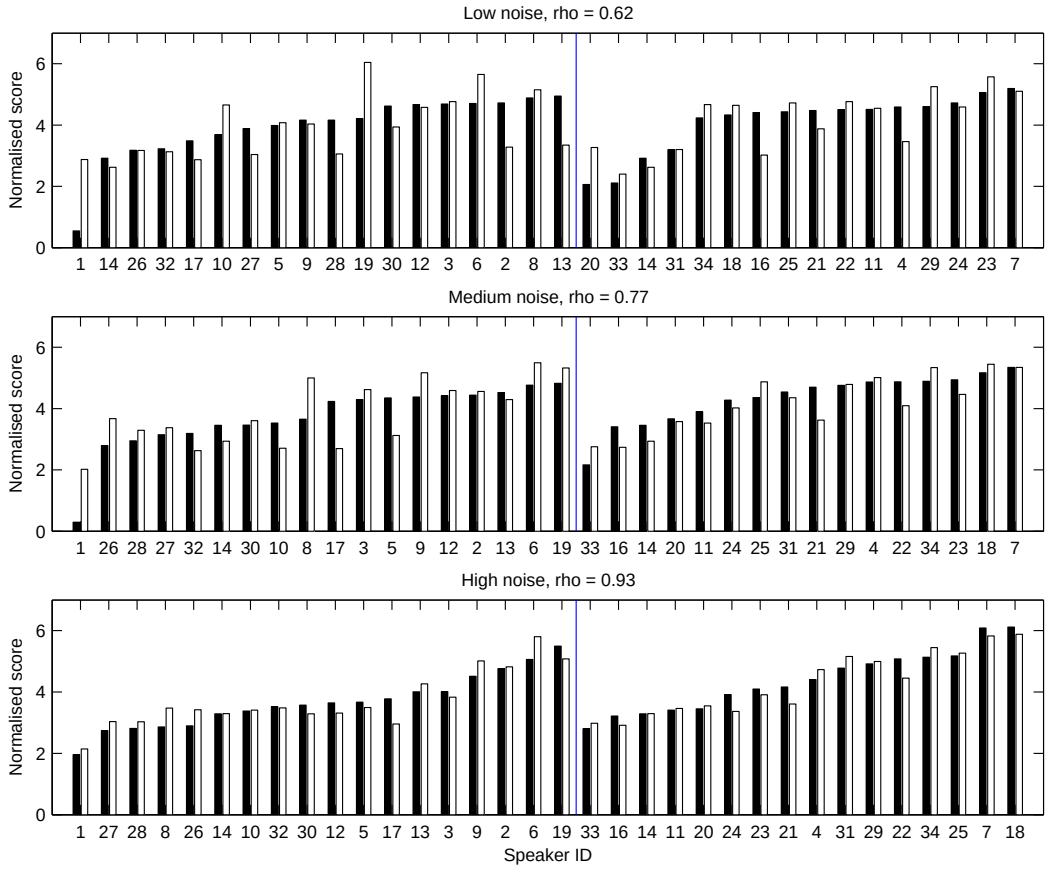

Fig. 12. *An illustration of the correlation between the ASR-based and HSR scores across all speakers in the Grid corpus for the low (top), medium (middle) and high (bottom) noise level bands. Male speakers are on the left of the chart and female speakers are on the right, with speakers ordered by increasing intelligibility (i.e. HSR score). Each pair of bars represents a separate speaker, with the bar on the left indicating the scaled HSR score and the bar to the right indicating the scaled ASR-based score. ASR-based and HSR scores have been normalised to have equal mean and variance before plotting so that the correlation is more readily apparent. The correlation coefficients are 0.61, 0.75 and 0.93 for the low, medium and high noise level bands respectively ($p<0.001$).*

One outcome was that female speakers were more intelligible than males in moderate and high noise levels, echoing the finding reported for clean speech in Hazan and Markham (2004). Further, individual speakers varied greatly in intelligibility. For example, in the high noise conditions, listeners identified 24% of keywords from the least intelligible speaker but scored 68% for the most intelligible speaker.

An across-utterance analysis of a number of acoustic parameters (VTL, mean F0 and duration/speaking rate) attempted to identify the basis for differences in intelligibility. Interestingly, while mean F0 and VTL were highly correlated when data from all speakers was pooled, no correlation was found within each

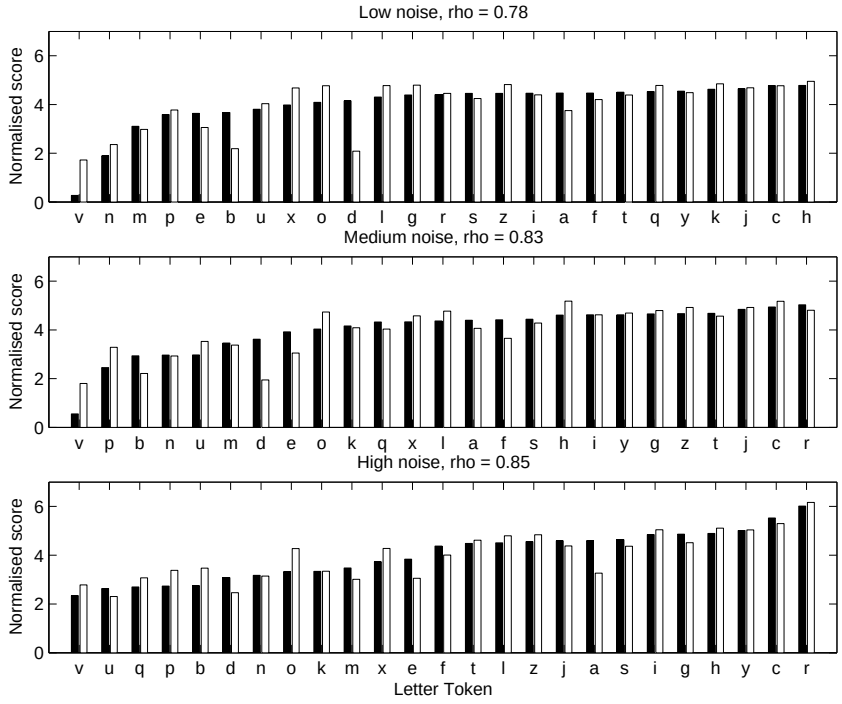

Fig. 13. *An illustration of the correlation between the ASR-based model and HSR scores across all letter tokens in the Grid corpus for the low (top), medium (middle) and high (bottom) noise level bands. Each pair of bars represents a separate letter token, with the bar on the left indicating the scaled HSR score and the bar to the right indicating the scaled ASR-based score. Tokens are order by increasing HSR score. ASR-based and HSR scores have been normalised to have equal mean and variance before plotting so that the correlation is more readily apparent. The correlation coefficients are 0.76, 0.82 and 0.87 for the low, medium and high noise level bands respectively ($p<0.001$).*

gender. Subsequently, male and female speakers were analysed separately. For females only, both VTL warp factor and mean F0 were positively correlated with intelligibility. For both groups, rapidly-spoken utterances were significantly less intelligible than those produced at a moderate rate, but the latter were no less intelligible than the slowest utterances.

However, although the 3 parameters showed some influence on the intelligibility of individual *utterances*, none of them were consistently good predictors of the intelligibility of individual *speakers*. This results suggests that within-speaker differences in mean F0 and duration are too large to act as a discriminative basis for speaker identification, and that differences in VTL across speakers are insufficient for fine-grained intelligibility prediction.

A simple measure of energetic masking was used to examine whether the 'visibility' of speech glimpses in noise could be used to predict speaker intelligibility. With utterances from all speakers pooled, visibility was highly-correlated with intelligibility both across SNRs and within a single SNR. Further, visibil-

ity was a good predictor of the intelligibility of individual females, especially in high noise conditions. However, visibility alone was not able to account for differences across the male speakers.

Recognition scores from a 'glimpsing' model which utilised visible spectro-temporal information and employed ASR techniques to train speaker-dependent statistical models were fitted to the behavioural data pooled across all speakers. Using the single set of parameters which resulted in the best overall fit, the model was able to predict not only the intelligibility of individual speakers to a remarkable degree (Figure 12), but could also account for most of the token-wise intelligibilities of the letter keywords (Figure 13). The fit was particularly good in high noise conditions.

6.2 Effect of noise level

In the low SNR condition, relative intelligibility estimates are a particularly good fit to listeners' data. This may be because many of the detailed phonetic cues and fine temporal structure that are poorly represented by the HMM-based acoustic models are equally inaccessible to humans when noise levels are high. In such conditions, listeners may have to resort to the same robust spectral envelope cues that the model employs. Model performance may also be improved at low SNRs because there is less redundancy in the separated glimpses of speech than there is in the complete spectro-temporal representation. The sparsity of the glimpses may better fit the independence assumptions of the HMM-based acoustic models. Indeed, it has been observed informally in previous missing data ASR studies that recognition can actually *improve* when a small amount of masking noise is added to clean speech.

The relative poorer performance of the model in the cleaner conditions may be a symptom of inadequacies of the acoustic modelling component. The frame-based spectral or cepstral HMMs employed by traditional ASR are crude representations of speech: they make invalid assumptions about the independence of adjacent frames; they do not model temporal fine structure; they poorly represent fine phonetic detail (Hawkins, 2003); and they are notoriously poor at modelling duration constraints. In recognition tasks which focus on acoustic modelling – i.e. those which require trivial language models – the best ASR systems fall well below the performance of humans (the performance gaps quoted by Lippmann (1997) have narrowed surprisingly little in the last 10 years). In the current task, poor acoustic modelling contributes to frequent confusions between acoustically similar token pairs such as 'b'/'v' and 'm'/'n' – confusions that listeners do not suffer from until a considerable amount of noise has been added to the speech. The model makes twice as many recognition errors as humans in low noise conditions. However, although this affects

absolute intelligibility, it does not necessarily have an impact on the prediction of *relative* speaker intelligibility. The model might consistently overestimate speaker errors by a fixed amount while getting the ordering of intelligibilities exactly correct. However, it is likely that acoustic modelling problems are speaker-specific and therefore introduce relatively more errors into some speakers than others. For example, there are problems that are specific to female speech: higher mean F0 values lead to unwanted resolution of harmonic structure which introduces a variability in the spectral profile.

6.3 Glimpse detection

The model simulates the detection of speech glimpses by using *a priori* knowledge of the unmixed signals and tunable local SNR (T) and minimum glimpse size (N) thresholds. Tuning these parameters revealed that the model fits the listener data when T is around 3 dB, and that there is a trade off between threshold and glimpse size. A similar fitting procedure was employed in Cooke (2006), who found close fits to listener performance at two different local SNR thresholds, one at around 8 dB (with a value sensitive to the glimpse size parameter, N) and a lower threshold in the range -5 to -2 dB. This earlier study differed from the current one in a number of ways. Most significantly, the current study used stationary noise at a range of SNRs, whereas Cooke (2006) employed noise with a varying degree of stationarity at a fixed global SNR of -6 dB. If the tuning of T is repeated in the current study using just the -6 dB global SNR data, good matches to listener data are found at -3 db and +3 dB, and the model-listener distance versus threshold curves look remarkably like those reported in Cooke's earlier study. Of course, it is unlikely that any 'glimpse detector' in the auditory system works at a fixed negative or positive local SNR threshold. Such mechanisms, if they exist, are likely to be far more complicated than the simple thresholding simulated here. For example, it is likely that periodic speech information can be detected at a lower local SNR than aperiodic speech information. Thresholds may be tunable – at higher global SNR the threshold may be lowered to allow through more (but noisier) data. Glimpse detection may even be under the guidance of top-down mechanisms, i.e. with more central recognition processes adapting the parameters of more peripheral processes such as to reduce recognition error. Considering the possible complexities, it is perhaps surprising that the simple model implemented here predicts relative speaker and token intelligibilities with such fidelity.

6.4 Informational masking

The current model provides no account of informational masking (IM). IM is a central process and is often discussed in terms of *attention*, i.e. the difficulty that the listener has in focusing on the masker and excluding the target. IM effects may also result from masker elements being incorrectly grouped with elements of the target. From a source-modelling perspective such as that offered by ASR, one might assume that IM results when glimpses of the masker fit well to the target model. In the current study, stationary speech-shaped noise was chosen as the masker in order to minimise IM effects – stationary noise provides a very effective *energetic masker*, but is not readily confusable with speech. However, it is possible that small glimpses of the noise background could be incorrectly assigned to the foreground. This would be mostly likely to occur when the speech and noise have the same average level (i.e. the 0 dB global SNR mixtures). As noted earlier, this IM-like effect is one possible explanation for the dip in correlation between estimated and actual intelligibilities seen at around 0 dB (see Figure 11).

6.5 Conclusions

Listeners identified keywords in short sentences spoken by a range of speakers presented in stationary noise at a number of SNRs. A model based on the recognition of glimpses of the target speech resulting from an energetic masking procedure provided good predictions of relative speaker intelligibility. The same model also predicted listeners' performance on individual keywords. A better fit was provided in conditions of moderate to high noise than in quiet and low noise, suggesting that while more detailed acoustic representations and models are required to capture the intrinsic intelligibility of individual speakers, relatively crude but robust representations may suffice when noise is present.

ACKNOWLEDGEMENTS

This study was supported by grants from the University of Sheffield Research Fund and the UK Engineering and Physical Research Council (grant GR/T04823/01).

References

W. A. Ainsworth and G.F. Meyer. Recognition of plosive syllables in noise: Comparison of an auditory model with human performance. *Journal of the*

Acoustical Society of America, 96:687–694, 1994.

P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA Proceedings*, volume 17, pages 97–110, 1993.

P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 4.3.14) [computer program], 2005. Retrieved May 26, 2005, from http://www.praat.org.

Z.S. Bond and T.J. Moore. A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, 14:325–337, 1994.

A.R. Bradlow, G.M. Torretta, and D.B. Pisoni. Intelligibility of normal speech. 1. global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20:255–272, 1996.

D. S. Brungart, B. D. Simpson, M. A. Ericson, and K. R. Scott. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America*, 110:2527–2538, 2001.

D.S. Brungart. Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, 109:1101–1109, 2001.

J.P. Burg. *Maximum entropy spectrum analysis*. PhD thesis, Stanford university, 1975.

D. van Compernolle. Recognizing speech of goats, wolves, sheep and ... nonnatives. *Speech Communication*, 35:71–79, 2001.

M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, 2001.

M. P. Cooke. Glimpsing speech. *Journal of Phonetics*, 31:579–584, 2003.

M.P. Cooke. A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, 199:1562–1573, 2006.

M.P. Cooke, J. Barker, S. P. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, submitted.

N.I. Durlach, C.R. Mason, G. Kidd Jr, T.L. Arbogast, H.S. Colburn, and B.G. Shinn-Cunningham. Note on informational masking. *Journal of the Acoustical Society of America*, 113:2984–2987, 2003.

H. Fletcher and R.H. Galt. The perception of speech and its relation to telephony. *Journal of the Acoustical Society of America*, 22:89–151, 1950.

O. Ghitza. Adequacy of auditory models to predict human internal representation of speech sounds. *Journal of the Acoustical Society of America*, 93: 2160–2171, 1993.

S. Hawkins. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31:373–405, 2003.

V. Hazan and D. Markham. Acoustic-phonetic correlates of talker intelligibility for adults and children. *Journal of the Acoustical Society of America*, 116(5):3108–3118, 2004.

I. Holube and B. Kollmeier. Speech intelligibility prediction in hearing-

impaired listeners based on a psychoacoustically motivated perception model. *Journal of the Acoustical Society of America*, 100:1703–1716, 1996.

J.-C. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, 93 (1):510–524, 1993.

J.C. Krause and L.D. Braida. Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America*, 115:362–378, 2004.

Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22:1–15, 1997.

H. Musch and S. Buus. Using statistical decision theory to predict intelligibility. i. model structure. *Journal of the Acoustical Society of America*, 109: 2896–2909, 2001.

M.A. Picheny, N.I. Durlach, and L.D. Braida. Speaking clearly for the hard of hearing. 1. intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, 28:96–103, 1985.

M.A. Picheny, N.I. Durlach, and L.D. Braida. Speaking clearly for the hard of hearing. 2. acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29:434–446, 1986.

R.V. Shannon, A. Jensvold, M. Padilla, M.E. Robert, and X. Wang. Consonant recordings for speech testing. *Journal of the Acoustical Society of America*, 106:L71–L74, 1999.

H.J.M. Steeneken and T. Houtgast. A physical method of measuring speech-transmission quality. *Journal of the Acoustical Society of America*, 67:318–326, 1980.

Q. van Summers, D.B. Pisoni, R.H. Bernacki, R.I. Pedlow, and M.A. Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, 84(3):917–928, 1988.

P.C. Woodland. Speaker adaptation for continuous density HMMs: A review. In *Proc. Adaptation-2001*, pages 11–19, 2001.