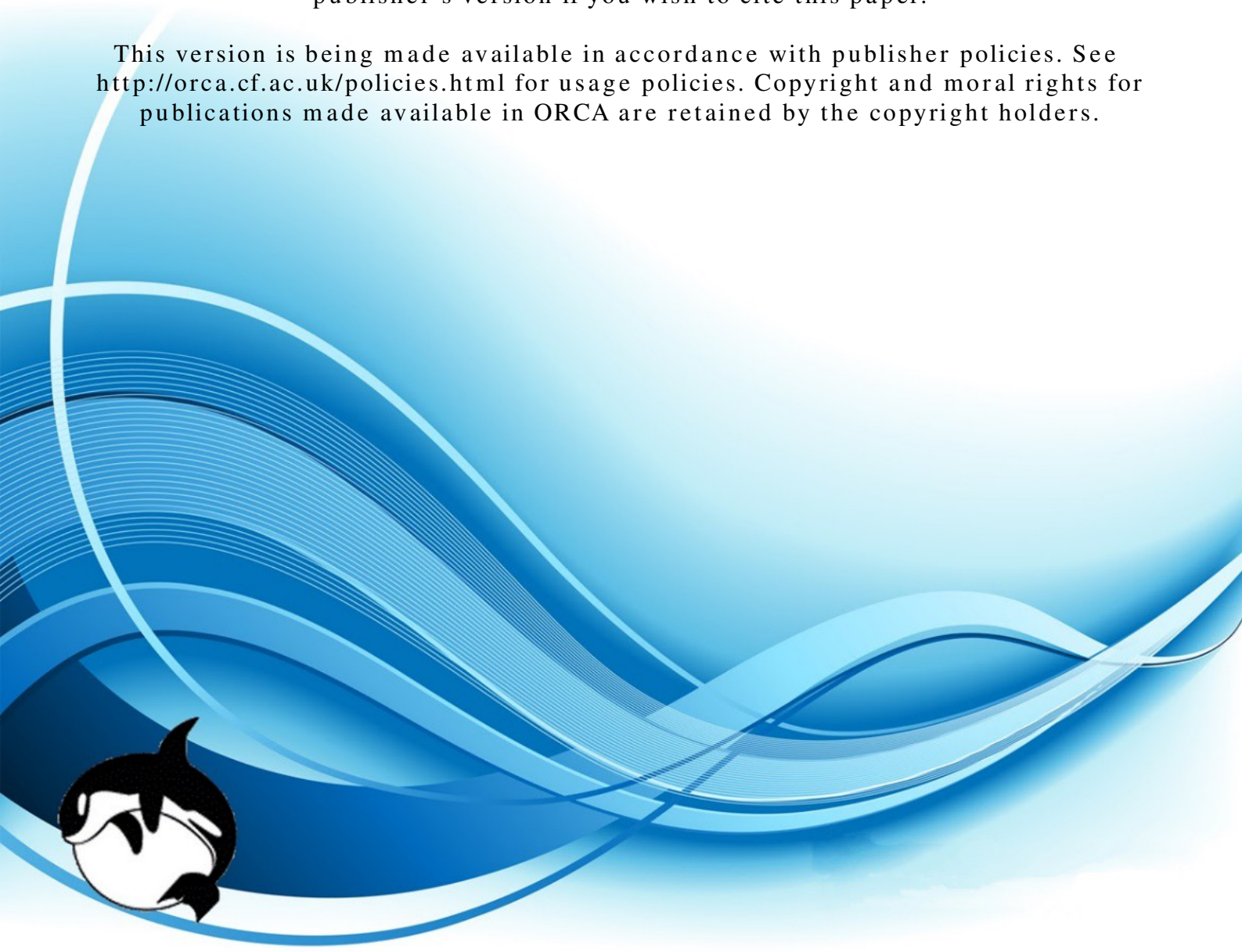

Speech perception from monaural and binaural information

John F. Culling, Barrie A. Edmonds, and Kathryn I. Hodder
School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT United Kingdom

(Received 13 September 2005; revised 7 October 2005; accepted 2 November 2005)

Two experiments explored the concept of the binaural spectrogram [Culling and Colburn, J. Acoust. Soc. Am. **107**, 517–527 (2000)] and its relationship to monaurally derived information. In each experiment, speech was added to noise at an adverse signal-to-noise ratio in the NoSπ binaural configuration. The resulting monaural and binaural cues were analyzed within an array of spectro-temporal bins and then these cues were resynthesized by modulating the intensity and/or interaural correlation of freshly generated noise. Experiment 1 measured the intelligibility of the resynthesized stimuli and compared them with the original NoSo and NoSπ stimuli at a fixed signal-to-noise ratio. While NoSπ stimuli were $\approx$50% intelligible, each cue in isolation produced similar (very low) intelligibility to the NoSo condition. The resynthesized combination produced $\approx$25% intelligibility. Modulation of interaural correlation below 1.2 kHz and of amplitude above 1.2 kHz was not as effective as their combination across all frequencies. Experiment 2 measured three-point psychometric functions in which the signal-to-noise ratio of the original NoSπ stimulus was increased in 3-dB steps from the level used in experiment 1. Modulation of interaural correlation alone proved to have a flat psychometric function. The functions for NoSπ and for combined monaural and binaural cues appeared similar in slope, but shifted horizontally. The results indicate that for sentence materials, neither fluctuations in interaural correlation nor in monaural intensity are sufficient to support speech recognition at signal-to-noise ratios where 50% intelligibility is achieved in the NoSπ configuration; listeners appear to synergistically combine monaural and binaural information in this task, to some extent within the same frequency region. [DOI: 10.1121/1.2140806]

PACS number(s): 43.71.An, 43.66.Pn, 43.66.Dc [PFA] Pages: 559–565

I. INTRODUCTION

When speech and interfering noise have different interaural time delays or phases, the speech is easier to understand than when they are presented with identical interaural parameters (Licklider, 1948; Schubert, 1956). The improvement in intelligibility is known as binaural advantage. It mirrors improvements in detection thresholds for tonal signals in similar binaural configurations known as the binaural masking level difference (BMLD). It would appear that the binaural system provides some additional information about the speech signal beyond that available by listening with one ear. However, while the BMLD has been intensively investigated, very little research has been directed to identifying the nature of binaurally derived speech information and the way in which it is combined with information derived monaurally.

Levitt and Rabiner (1967b) developed a technique for predicting the magnitude of the binaural advantage for speech. This technique assumed that the binaural system reduced the effective level of the masking noise in each frequency band in accordance with the BMLD of a pure tone at the same frequency and in the same binaural configuration; if the BMLD was 5 dB, then the effective masker level was reduced by 5 dB. For a given binaural configuration, the model used measurements of the pure-tone BMLD at each frequency and the articulation index (Kryter, 1962) to predict the SRT that would result from using that binaural configuration with speech signals. This model produced accurate predictions of their speech intelligibility measurements (Levitt and Rabiner, 1967a). However, the model is rather agnostic about the level of processing at which listeners combine monaural and binaural information; it simply assumes an effective equivalence between binaural unmasking and reduced monaural masking, which is, as they noted, "reasonable…at an empirical level." It would be interesting to determine whether monaural and binaural information is combined before or after word recognition.

Akeroyd and Summerfield (2000) addressed the combination of monaural and binaural cues in the specific context of vowel identification from formant information. In their experiments, formants were represented either by prominent frequency bands in the amplitude spectrum of a broadband noise or by interaurally decorrelated subbands of that noise. Akeroyd and Summerfield demonstrated that two formants encoded in these different forms could be combined to produce accurate identification of the vowel they represent, where information from either cue alone would leave the vowel ambiguous. Akeroyd and Summerfield's demonstration begins to explore the nature of the cues provided by each system and proves that information derived independently from each cue can be integrated. Although the stimuli employed were somewhat schematic, they reflect a situation that probably arises frequently in everyday listening, where binaural information is recovered effectively at low frequencies, but not at high frequencies. Vowel identification may therefore often rely upon a first formant frequency derived from binaural cues and a second formant frequency that is better represented by monaural cues. The present investiga-

tion looks at the same cues as Akeroyd and Summerfield, but using more naturalistic stimuli, derived from sentences embedded in noise, and also tests, to some degree, the integration of these cues within the same frequency region, as well as across frequency. The cues involved are temporal modulations within individual frequency channels in the amplitude and in the interaural coherence of the stimulus.

One consequence of adding a signal to a more-intense noise with a different interaural time delay is that the interaural coherence of the stimulus (the maximum interaural correlation[1] as a function of delay) is reduced at the frequency of the signal. As the intensity of the signal is increased the interaural coherence of the stimulus is reduced. In the widely used NoSπ binaural configuration, where the noise has zero interaural delay, the interaural coherence and the interaural correlation will be identical for negative signal-to-noise ratios. Consequently, it is often sufficient to consider only the interaural correlation (without applying any delays). Supporting the notion that interaural coherence/correlation is the cue used by the binaural system to produce binaural advantage, Culling *et al.* (2001) showed that direct manipulation of the interaural correlation of one subband of an otherwise diotic broadband noise generates the illusion of an embedded sound at the frequency of this subband. Further, they found that the lower the interaural correlation, the louder is the illusory sound. This experiment shows that the auditory system interprets reduced correlation as evidence of an additional sound. The perceived loudness increases as correlation is reduced, which is consistent with the expected effect of an increase in the level of that sound on the correlation.

In order to explain binaural advantage for complex, time-varying signals, such as speech, Culling and Colburn (2000) proposed a conception of binaural processing, in which spectro-temporal variations in signal intensity are recovered from the combined speech and noise by analyzing the interaural coherence of the stimulus within a sliding temporal window for each frequency channel. Based on measurements of binaural frequency selectivity (Kohlrausch, 1988) and temporal acuity (Culling and Summerfield, 1998), the frequency channels were modeled by a gammatone filterbank (Patterson *et al.*, 1987, 1988) followed by a sliding analysis window of asymmetric, Gaussian shape, and 110 -ms duration. In Culling and Colburn (2000), this analysis was followed by an interaural coherence measurement, giving a physical measure of cue strength. However, a further refinement to the model is to perceptually scale the interaural coherence values according to the sensitivity of the binaural system. Such a scaling provides the basis on which the brain encodes the intensity of the stimuli it detects using the binaural system. Culling *et al.* (2001) measured listeners' sensitivity function using cumulative d' to the interaural correlation of one subband in an otherwise diotic noise. Culling (2000) employed this sensitivity function to predict the salience of the Fourcin pitch as a function of the size of interaural delays and the number of noises used to generate the pitch. Thus, the output of the model is a salience measure that can be used to address suprathreshold as well as threshold data.

In the present investigation, the purpose was not to model human binaural processing, but to isolate the cues to which this model is sensitive and investigate listeners' ability to use such cues for speech recognition. In this way, we seek to underpin the claim that the model provides a plausible account of human proficiency at understanding speech in noise in the NoSπ configuration. In particular, the model assumes that the only information needed from the binaural system is the coherence, and not, for instance, the location of the speech or of the noise. If the spectro-temporal pattern of coherence is simulated and listeners can exploit the simulated binaural cues successfully, the contention that coherence is the only important cue receives some support. For this purpose, a method was developed for the analysis and resynthesis of monaural and binaural cues.

II. EXPERIMENT 1

Experiment 1 measured the intelligibility of speech in six different conditions. Two of these were real speech-and-noise mixtures in the NoSo and NoSπ binaural configurations. Three more were based upon the NoSπ stimuli and generated by analyzing and resynthesizing specific cues using the method outlined below. These conditions were (1) "binaural" using stimuli with resynthesized binaural cues, (2) "monaural" using stimuli with resynthesized monaural cues, and (3) "combined" using stimuli with both resynthesized binaural and monaural cues. Since effective binaural processing of speech in noise is known to be largely limited to low frequencies (Levitt and Rabiner, 1967a, b), a sixth "split" condition explored the possibility that binaural cues operate mainly below 1.2 kHz, while monaural cues operate mainly above 1.2 kHz. This condition was identical to the binaural stimuli below this frequency and identical to the monaural stimuli above it.

A. Analysis and resynthesis method

The general analysis and resynthesis method began with an input speech waveform (16-bit quantization; 20-kHz sampling rate) and a specified signal-to-noise ratio (SNR). Brown noise (i.e., with a −6-dB/oct roll-off) was generated and added to the speech in the NoSo and NoSπ binaural configurations at the desired SNR. The resulting stereo waveforms were both stored and the NoSπ waveform was analyzed as follows.

The left- and right-ear channels were each filtered in the frequency domain into 30 spectrally contiguous, logarithmically spaced frequency bands between 0.1 and 10 kHz. To filter out each band, the complete waveform was transformed into the frequency domain and all frequency bins outside the passband were set to zero amplitude before inverse Fourier transformation. The frequency resolution of this analysis was roughly equivalent to human frequency selectivity (Moore and Glasberg, 1983), but using rectangular passbands. The time waveform from each resulting frequency band was recovered by inverse Fourier transform and then windowed using a series of 50% overlapping Hanning windows of 100-ms equivalent-rectangular duration (200 ms in total duration). Two statistics were measured from the resulting

spectro-temporal bins: (1) the rms power of each bin for the left channel and (2) the correlation coefficient (Pearson's r) for corresponding bins from the left- and right-hand channels.

The stimuli of the monaural, binaural, and combined conditions contained only noise. These noises were manipulated in order to introduce the speech cues that had been determined using the analysis above. The cues were introduced as follows. Two freshly generated, independent Brown noises of equal duration to the original speech waveform were filtered and windowed in a similar fashion to the speech-noise mixture in order to produce two spectro-temporal arrays of noise bins. The measured interaural correlations and rms powers of the corresponding bins in the speech-noise mixture were imposed on the noise bins. Corresponding bins from the two noises were orthogonalized using the Gram-Schmidt method (Culling *et al.*, 2001, Appendix I), so that they had an interaural correlation of zero and equal rms power. To recreate the original pattern of interaural correlation changes (binaural condition), each pair of noise bins was then mixed[2] so that their interaural correlation exactly matched that of the corresponding pair of spectro-temporal bins from the speech-noise mixture. To recreate the original power spectrogram (monaural condition), the rms power of each noise bin was measured and then the sample values were scaled so as to match the power of the corresponding spectro-temporal bin from the speech-noise mixture. Both of these operations were independent, so each speech-noise mixture stimulus was resynthesized from noise alone with monaural cues, binaural cues, or both.

The binaural cues alone result in the perception of muffled speech embedded in the background noise. When the content of the sentence is known, the content of the sentence can also be "followed," but, as the experiments below demonstrate, without prior knowledge, it is very difficult to understand a novel speech sample from binaural cues alone.

B. Stimuli

The stimuli were based on 60 sentences from the Harvard sentence list (M.I.T. recordings of voice DA), which are equalized in overall level. Each sentence was processed into the six different forms described above using an original SNR of −20 dB. A further ten sentences were prepared for use as practice stimuli in the NoSo and NoSπ binaural configurations. The SNRs for these practice stimuli ranged from −12 to −20 dB in 2-dB steps. The presentation level was approximately 57 dB (A).

C. Procedure

Twelve listeners with no reported hearing defects and all native speakers of English each attended a single hour-long session, during which they listened to 70 sentences. The first ten sentences were a practice, which was identical for each listener. The ten practice sentences alternated between the NoSo and NoSπ binaural configurations and decreased in SNR by 2 dB on every other trial, such that the SNR matched that of the test stimuli by the end of the practice. In the subsequent 60 trials listeners received ten sentences in each of the six conditions in randomized sequence. The stimulus materials were rotated through these conditions from one listener to the next, so that each sentence was represented equally in each of the conditions.

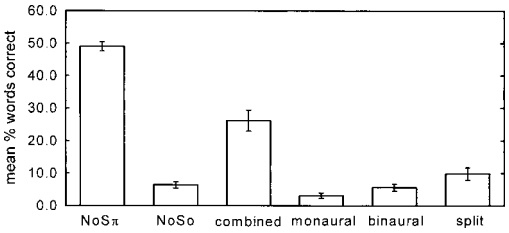

FIG. 1. Results from experiment 1. Percentage of words correctly recognized in six conditions: NoSπ, NoSo, "combined," (resynthesized combination of the monaural and binaural cues of amplitude and correlation modulation), "monaural" (resynthesized monaural cues only), "binaural" (resynthesized binaural cues only), and "split" (resynthesized binaural cues below 1.2 kHz and monaural cues above 1.2 kHz). Error bars are one standard error of the mean.

Listeners were seated in a single-walled IAC booth within a sound-treated room and made responses on a computer terminal, whose keyboard was within the booth. The screen could be seen through the booth's window. Listeners were invited to replay each stimulus (by pressing the "?" key) up to ten times before attempting a transcription. However, they were advised that no further benefit should be expected after listening three or four times to each stimulus. Once they had listened enough, they typed in their transcription and pressed the return key. The correct transcript then appeared on the screen, beneath their own, and they counted the number of words correctly transcribed and indicated the total with a single keypress, 0–9. Five of the sentences contained ten words, but it was very rare for all ten to be correctly transcribed. In this situation, listeners were instructed to press "9."

D. Results

The mean percentage of words correctly recognized in each condition is shown in Fig. 1. These are averages across listeners and represent the total number of words recognized from the ten sentences presented in a given condition. Since the sentences were not all of equal length, some variance arises from differences in the maximum possible score for a given listener in a given condition, but this effect is counterbalanced across conditions by the rotation of materials, as are any variations in the intrinsic intelligibility of different sentence lists. An analysis of variance showed that the six conditions differed significantly $[F(5,55)=119,\ p<0.001]$. However, Tukey HSD pairwise comparisons indicated that the NoSo, monaural, binaural, and split conditions did not differ significantly, while the NoSπ and combined conditions differed from these and from each other ($q>10$, $p<0.01$, in each case). The NoSπ condition is therefore significantly better than any other. The combined condition displayed a marked superiority over the monaural and binaural conditions, but was still worse than the NoSπ condition, from whose stimuli it was derived.

E. Discussion

The low recognition rate in the binaural condition of experiment 1 indicates that binaural cues alone are insufficient to facilitate substantial intelligibility. Indeed many of the correctly recognized words probably occurred by chance; the words "a" and "the" accounts for about 50% of correct words in the more difficult conditions. Nonetheless, some transcripts show unambiguous evidence of intelligibility, supporting the principle that the binaural spectrogram can support a low, but measurable, level of speech communication. Of 120 attempted transcripts from the 12 subjects, the most accurate was "we all have a new day's decision" for "we now have a new base for shipping." Aside from the four correct words, there is a matching prosodic structure, the correct number of syllables, and many phonemes within incorrect words either match or have the correct manner of articulation.

Although both monaural and binaural cues alone gave poor intelligibility, it is striking that their combination resulted in a large increase in scores. For instance, in the combined condition, "raise the sail and steer the ship northward" was transcribed quite accurately by one listener as "raise the sail, steer the ship upward." The proposed direction of movement suggests that this listener was not, while making these transcriptions, making very great use of semantic constraints. The success of the combined condition suggests two possibilities. One is that there is some form of perceptual interaction between the two types of cue, whereby one cue corroborates or reinforces the other. The other is that there is simply a highly nonlinear relationship between the information gained about the stimulus and the chance of correctly identifying a whole word. Experiment 2 was designed to address this question.

Although speech recognition was markedly improved by the combination of monaural and binaural cues in the combined condition it still did not match that of a real mixture of speech and noise in the NoSπ condition. When frequency bands above and below 1.2 kHz were encoded with different perceptual cues (split condition), performance was not significantly better than when the individual cues were provided at all frequencies (monaural and binaural conditions). This outcome suggests that monaural and binaural cues are exploited to some extent from the same frequency regions. Since it is known from the work of Levitt and Rabiner (1967a, b) that the contribution of the binaural system at high frequencies is very limited, one must conclude that substantial monaural information is normally extracted from the band below 1.2 kHz, and that this information remains useful to speech intelligibility when corresponding binaural information is also available.

III. EXPERIMENT 2

In experiment 1, listeners showed markedly higher identification rates when provided with both monaural and binaural cues, compared to either cue in isolation. If information from monaural and binaural cues is independent and is combined at the lexical level, then the size of this improvement should conform to an equation offered by Boothroyd and Nittrouer (1988), which sums log error rates:

$$\log(1-p_c)=\log(1-p_m)+\log(1-p_b). \quad (1)$$

The equation is adapted so that, p_m, p_b, and p_c are the proportions of words correct in the monaural, binaural, and combined conditions, respectively. One implication of this equation is that when two independent sources of information are combined, the resulting psychometric function is predicted to be steeper than that produced when either cue is available individually. If this equation is applied to the results of experiment 1, the predicted score for the combined condition is only 9%, compared to the observed 27%. However, the method of scoring based on all the words is probably not best adapted to this form of analysis, since there are substantial opportunities to get common words such as "the" right by chance. Therefore, experiment 2 employed a scoring method that used a fixed number of content words. In addition, the signal-to-noise ratio of the original NoSπ stimulus was varied, so that a three-point psychometric could be measured for each condition.

A. Stimuli

The stimuli were generated in a similar manner to those of experiment 1, but in order to accommodate testing at three SNRs, the number of conditions was reduced to four, NoSπ, monaural cues only, binaural cues only, and combined monaural and binaural cues. In addition, the quantity of starting speech material had to be increased to 120 sentences for experimental materials and another 20 for practice. The three SNRs for the original speech-noise mixture were −20, −17, and −14 dB. SNR thus increased in 3-dB steps from that used for experiment 1. With four binaural conditions and three SNRs, there were 12 conditions in all. The presentation level was approximately 57 dB (A).

B. Procedure

Sixteen listeners each attended one 75-min session. During a session, they listened to 20 practice stimuli including five examples from all four conditions at declining SNR as in experiment 1. They then listened to 120 experimental stimuli (10 examples $\times$ 4 conditions $\times$ 3 SNRs). The 120 stimuli were divided into three blocks. Each successive block used stimuli based on a lower SNR. As in experiment 1, each listener heard ten sentences in each condition, and the sentences were rotated around the conditions from one listener to the next. Unlike experiment 1, listeners heard each sentence three times and scoring was based upon five keywords, indicated to the listeners with the use of capital letters in the transcript.

C. Results

The results are plotted as psychometric functions in Fig. 2. The abscissa represents the SNR of the NoSπ stimuli and the SNR of the NoSπ stimuli on which the other three conditions were based. The four stimulus types differed significantly $[F(3,45)=340,\ p<0.0001]$. The NoSπ condition showed significantly higher intelligibility than all other con-

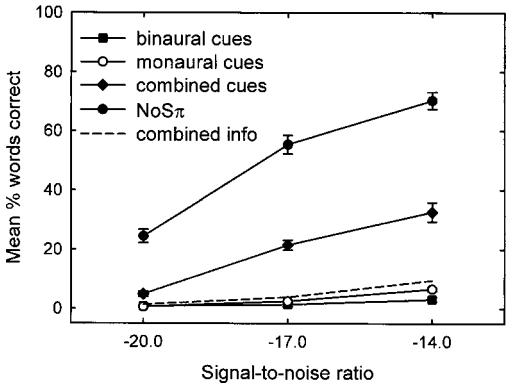

FIG. 2. Results from experiment 2. Percentage of words correctly recognized as a function of signal-to-noise ratio in four conditions: NoSπ (filled circles), resynthesized binaural cues only (filled squares), resynthesized monaural cues (open circles), resynthesized combination of monaural and binaural cues (filled diamonds), and expected scores from combined monaural and binaural cues (dashed line). Error bars are one standard error of the mean.

ditions ($q>24$, $p<0.001$, in each case). The combined cues also gave significantly higher intelligibility than either monaural or binaural cues ($q>13$, $p<0.001$, in each case).

Two of the four conditions showed a substantial improvement in intelligibility with increasing SNR, but the monaural-cue and binaural-cue conditions showed little or no improvement. This is reflected in a significant main effect of SNR [$F(2,30)=188$, $p<0.0001$], but an interaction between SNR and condition [$F(6,90)=55.0$, $p<0.0001$]. Simple main effects showed significant improvement with SNR for the NoSπ condition [$F(2,30)>240$, $p<0.0001$], combined-cue condition [$F(2,30)>86$, $p<0.0001$], and a small but significant improvement in the monaural-cue condition [$F(2,30)>4.2$, $p<0.05$], but not for the binaural-cue condition.

Predictions of performance in the combined condition based on that from the monaural and binaural conditions using Eq. (1) are shown by a dashed line in Fig. 2. These predictions are substantially lower than the observed level of performance in the combined condition.

D. Discussion

The results of experiment 2 indicate not only that isolated binaural cues cannot support full speech intelligibility, but that there is no sign of any improvement with increasing SNR. Even the monaural-cue condition shows only the slightest of improvements across this range of SNRs. The substantial superiority of the combined cue condition over both the binaural-cue and monaural-cue conditions suggests that the information furnished by these two cues is in some way complementary. The fact that the predictions based on Eq. (1) fall well below the observations in the combined condition indicates that information from the two types of cue is not combined at the lexical level and probably occurs at a lower level of processing. The exact nature of that complementarity remains elusive.

Because the individual-cue psychometric functions are so flat, it is not possible to make a comparison of slopes. However, the results clarify and amplify those of experiment 1. The combined cues produced growing intelligibility with SNR, while the corresponding individual cues supported only the slightest improvement. A remarkable feature of the data from experiment 2 is that the intelligibility furnished by the binaural system did not seem to grow at all with increasing SNR, so it may be the case while that the binaural system provides some information that contributes to intelligibility, it rarely provides sufficient information to support the accurate identification of words. It is tempting to speculate that this information is fairly constant in usefulness over a wide range of SNRs and the shape of the psychometric function for combined cues is therefore determined only by the increasing value of monaural cues.

Finally, it is noteworthy that at the most adverse SNR used in experiment 2, listeners did not recognize as many words as in experiment 1, which used the equivalent SNR. This difference may have arisen from three sources. First, in experiment 2 these stimuli were presented within the context of an experiment that included much more favourable SNRs and this may have reduced the level of effort that the listeners were prepared to devote to the more difficult stimuli. Second, experiment 2 used a different scoring system based on keywords correct, rather than all correct words. This may have reduced scores on some of the more difficult conditions, for which articles accounted for many of the correct words (in the monaural and binaural conditions, correct words were 51% articles, 15% verbs, 5% personal pronouns, 10% adjectives, 9% prepositions, and 10% nouns). Third, listeners only had three opportunities to repeat the stimulus in experiment 2 compared to up to ten in experiment 1.

IV. GENERAL DISCUSSION

The experiments in this study have explored the plausibility of a theory of binaural unmasking for speech put forward by Culling and Colburn (2000). This theory suggested that listeners detect spectro-temporal variations in interaural coherence and treat these as a perceptual surrogate for the spectro-temporal variations in the energy of the speech signal. The results offer support for the theory in that (1) stimuli designed to provide listeners with this cue in isolation facilitated measurable, if very low, intelligibility, and (2) when these cues were combined with monaural cues, a substantial increase in intelligibility was observed, mirroring the effect of binaural advantage. This said, it is disappointing that the isolated binaural cues did not facilitate greater speech recognition, and that the resynthesis of combined monaural and binaural cues did not fully emulate the intelligibility of speech in noise in the NoSπ configuration. It is not clear whether these failings arise from a limitation of the theory or of the signal processing.

It is possible that there was some lack of fidelity in reproducing the pattern of amplitude and interaural correlation changes from the original stimulus and that this introduced informational noise to the cues the listeners were using. An obvious possibility here is that analysis windows of 100 ms, while being appropriate for the binaural cues, are very long compared to the 10-ms temporal resolution of the monaural system (Plack and Moore, 1990). Another is that

the Fourier-based method of filtering produced excessive ringing. Alternatively there may be some additional cue, perhaps provided by the fine temporal structure of the stimulus that listeners lacked when they were not listening to a real speech-noise mixture. Some attempts were made to address these possibilities, but no clear leading candidate emerged. First, analysis within shorter windows did not yield an audible improvement (not formally tested). Second, visual inspection of spectrograms of the frequency bands produced by rectangular-passband, frequency-domain filtering did not reveal ringing that seemed likely to be significant at the 100-ms time scale on which the processing was based. Third, a simple test of the integrity of the stimuli was to subtract one channel from the other. For NoSπ stimuli, this will recover the original speech waveform, but for the binaural cues and combined cues conditions, it results in a spectro-temporally modulated noise similar to those generated by Shannon *et al.* (1995), but with a modulation spectrum that is low-pass filtered at just 10 Hz. These stimuli sounded like an individual with some serious glottal pathology, but were nonetheless highly intelligible. It is possible that fine temporal structure makes a difference. An obvious possibility is that the presence of the target talker's fundamental frequency allows the auditory system to select the harmonics of that fundamental frequency. Although a persuasive idea, it runs against current theories of fundamental frequency processing that suggest a process of harmonic cancellation, rather than selection (de Cheveigné, 1997).

The reliance of Culling and Colburn's theory on measurement of interaural coherence implicitly limits its scope to speech recognition in unmodulated noise. This is because the direct relationship between interaural coherence and signal intensity is disturbed if the masker is modulated. For present purposes, it is convenient to remain within this constraint, because it facilitates the form of resynthesis used here. However, a better theory would more successfully predict performance with modulated maskers, such as competing speech. van de Par *et al.* (2001) argued that subtractive mechanisms such as Durlach's (1963) E-C model are more robust in modulated maskers. An improved form, therefore, of Culling and Colburn's model might employ a mechanism similar to Culling and Summerfield's (1995) mE-C model within each spectro-temporal bin. Such a model would equalize the level and interaural delay independently within each spectro-temporal bin and return the residue from cancellation as a measure of the signal intensity.

ACKNOWLEDGMENTS

The authors would like the Hanse Wissenschaftskolleg, which supported Culling with a fellowship during the course of this study, the editor, Peter Assmann, and the two reviewers, Chris Darwin and Wes Grantham, for their valuable comments on the manuscript.

[1]Throughout the present article "interaural correlation" may be taken to mean the Pearson's product-moment correlation, r (also known as the normalized covariance), of the waveforms at the two ears. At high frequencies and particularly for narrow stimulus bandwidths, the adopted definition of correlation can make a material difference to predictive models and the normalized interaural correlation has been found to facilitate better fitting models of interaural discrimination data than the normalized covariance (Bernstein, and Trahiotis, 1992, 1996a, b; Bernstein *et al.* 1999). However, for the broadband, mainly low-frequency stimuli used here, it makes little difference, so Pearson's r is employed here, as it is the more familiar.

[2]Mixing was performed using the "two-noise" method, in which one channel contains noise N_1 and the other mixture of the two noises, N_1 and N_2, in the following proportions: $\rho N_1 + \sqrt{1-\rho^2} N_2$.

Akeroyd, M. A. and Summerfield, A. Q. (**2000**). "Integration of monaural and binaural evidence of vowel formants," J. Acoust. Soc. Am. **107**, 3394–3406.

Bernstein, L. R., and Trahiotis, C. (**1992**). "Measurements of interaural envelope correlation and its relation to binaural unmasking at high frequencies," J. Acoust. Soc. Am. **91**, 306–316.

Bernstein, L. R., and Trahiotis, C. (**1996a**). "On the use of the normalized correlation as an index of interaural envelope correlation," J. Acoust. Soc. Am. **100**, 1754–1763.

Bernstein, L. R., and Trahoitis, C. (**1996b**). "The normalized correlation: Accounting for binaural detection across center frequency," J. Acoust. Soc. Am. **100**, 3774–3784.

Bernstein, L. R., van de Par, S., and Trahiotis, C. (**1999**). "The normalized correlation: Accounting for NoS thresholds obtained with Gaussian and 'low-noise' masking noise," J. Acoust. Soc. Am. **106**, 870–876.

Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**, 101–114.

Culling, J. F. (**2000**). "Dichotic pitches as illusions of binaural unmasking. III. The existence region of the Fourcin pitch," J. Acoust. Soc. Am. **107**, 2201–2208.

Culling, J. F., and Colburn, H. S. (**2000**). "Binaural sluggishness in the perception of tone sequences and speech in noise," J. Acoust. Soc. Am. **107**, 517–527.

Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98**, 785–797.

Culling, J. F., and Summerfield, Q. (**1998**). "Measurements of the binaural temporal window using a detection task," J. Acoust. Soc. Am. **103**, 3540–3553.

Culling, J. F., Colburn, H. S., and Spurchise, M. (**2001**). "Interaural correlation sensitivity," J. Acoust. Soc. Am. **110**, 1020–1029.

de Cheveigné, A. (**1997**). "Concurrent vowel identification III: A neural model of harmonics interference cancellation," J. Acoust. Soc. Am. **101**, 2857–2865.

Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking level differences," J. Acoust. Soc. Am. **35**, 1206–1218.

Kohlrausch, A. (**1988**). "Auditory filter shape derived from binaural unmasking experiments," J. Acoust. Soc. Am. **84**, 573–583.

Kryter, K. (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1684–1697.

Levitt, H., and Rabiner, L. R. (**1967a**). "Binaural release of masking for speech and gain in intelligibility," J. Acoust. Soc. Am. **42**, 601–608.

Levitt, H., and Rabiner, L. R. (**1967b**). "Predicting binaural gain in intelligibility and release from masking for speech," J. Acoust. Soc. Am. **42**, 820–828.

Licklider, J. C. R. (**1948**). "The influence of interaural phase relations on the masking of speech by white noise," J. Acoust. Soc. Am. **20**, 150–159.

Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987). "An efficient auditory filterbank based on the gammatone function," paper presented to the IOC speech group on auditory modeling at RSRE, 14–15 December.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1988). "Spiral VOS final report, Part A: The auditory filter bank," Cambridge Electronic Design, Contract Report (APU 2341).

Plack, C. J. and Moore, B. C. J. (**1990**). "Temporal window shape as a function of frequency and level," J. Acoust. Soc. Am. **87**, 2178–2187.

Schubert, E. D. (**1956**). "Some preliminary experiments on binaural time delay and speech intelligibility," J. Acoust. Soc. Am. **28**, 895–901.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekied, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
van de Par, S., Trahiotis, C., and Bernstein, L. R. (**2001**). "A consideration of the normalization that is typically included in correlation-based models of binaural detection," J. Acoust. Soc. Am. **109**, 830–833.