Binaural sluggishness in the perception of tone sequences and speech in noise

John F. Culling and H. Steven Colburn
Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, Massachusetts 02215

(Received 2 November 1998; revised 19 July 1999; accepted 28 September 1999)

The binaural system is well-known for its sluggish response to changes in the interaural parameters to which it is sensitive. Theories of binaural unmasking have suggested that detection of signals in noise is mediated by detection of differences in interaural correlation. If these theories are correct, improvements in the intelligibility of speech in favorable binaural conditions is most likely mediated by spectro-temporal variations in interaural correlation of the stimulus which mirror the spectro-temporal amplitude modulations of the speech. However, binaural sluggishness should limit the temporal resolution of the representation of speech recovered by this means. The present study tested this prediction in two ways. First, listeners’ masked discrimination thresholds for ascending vs descending pure-tone arpeggios were measured as a function of rate of frequency change in the NoSo and NoSπ binaural configurations. Three-tone arpeggios were presented repeatedly and continuously for 1.6 s, masked by a 1.6-s burst of noise. In a two-interval task, listeners determined the interval in which the arpeggios were ascending. The results showed a binaural advantage of 12–14 dB for NoSπ at 3.3 arpeggios per s (arp/s), which reduced to 3–5 dB at 10.4 arp/s. This outcome confirmed that the discrimination of spectro-temporal patterns in noise is susceptible to the effects of binaural sluggishness. Second, listeners’ masked speech-reception thresholds were measured in speech-shaped noise using speech which was 1, 1.5, and 2 times the original articulation rate. The articulation rate was increased using a phase-vocoder technique which increased all the modulation frequencies in the speech without altering its pitch. Speech-reception thresholds were, on average, 5.2 dB lower for the NoSπ than for the NoSo configuration, at the original articulation rate. This binaural masking release was reduced to 2.8 dB when the articulation rate was doubled, but the most notable effect was a 6–8 dB increase in thresholds with articulation rate for both configurations. These results suggest that higher modulation frequencies in masked signals cannot be temporally resolved by the binaural system, but that the useful modulation frequencies in speech are sufficiently low (<5 Hz) that they are invulnerable to the effects of binaural sluggishness, even at elevated articulation rates. © 2000 Acoustical Society of America. [S0001-4966(00)02601-1]

PACS numbers: 43.66.Pn, 43.66.Ba, 43.66.Dc, 43.66.Hg [DWG]

INTRODUCTION

This investigation combines two recent concepts in binaural research and applies them to the practical issue of how the binaural system assists the understanding of speech in noise. The first concept is the theory that when signals are added to noise at a differing interaural phase or delay, they are detected by virtue of the changes in interaural correlation, which target signals induce in noise (Durlach et al., 1986); the second is the sluggishness of the binaural system, which seems unable to follow rapid changes in the temporal relationships between the sound reaching the two ears, as though this information is temporally smeared within the auditory system (Grantham and Wightman, 1978, 1979). Taken together, these ideas imply first, that the modulations of a speech signal within different frequency bands are detected by the binaural system as modulations in the interaural correlation of those frequency bands, and second, that these modulations in correlation are perceived only at low modulation rates. For the situations encountered within this study, addition of the signal always causes a reduction in interaural correlation.

It will be demonstrated below that, according to this logic, the binaural system should lose the spectrotemporal information which codes formant movements, while preserving information associated with the enunciation of syllables. One might expect the loss of formant-movement information to reduce the usefulness of the binaural system in speech understanding. The effectiveness of the binaural system for speech recognition can be measured by comparing the masked speech-reception thresholds (SRTs) in the NoSo and NoSπ configurations. The difference in threshold is called the binaural intelligibility level difference (BILD). However, although the BILD is only 5–6 dB, whereas masked pure-tone thresholds can show differences of 15 dB, this difference in the size of the masking release can be fully accounted for without reference to binaural sluggishness.

Levitt and Rabiner (1967a,b) showed that the smaller size of the BILD compared to binaural masking level differ-
ences (BMLDs) can be fully explained by the limited range of frequencies over which the binaural system is effective. They predicted a BILD of 5-6 dB from pure-tone data by using Fletcher’s articulation index (Fletcher and Galt, 1950; Kryter, 1962a,b). The articulation index, originally developed for the telecommunications industry, predicts the intelligibility of speech through a given transmission channel using information about that channel’s frequency response and noise; Levitt and Rabiner’s calculations assumed that the effective noise level at each frequency was reduced in favorable binaural configurations in accordance with the corresponding pure-tone BMLD at that frequency. In terms of the structural features of speech, the BILD is only 5–6 dB, because only the first formant region is strongly assisted by the binaural system. The articulation index does not, however, take into account the channel’s modulation transfer function (MTF), a measure of the fidelity with which a channel transmits different modulation frequencies. The MTF was developed for the purpose of predicting speech intelligibility in reverberant auditoria which produce temporal smearing of the signal (Houtgast and Steeneken, 1985). Since the binaural system seems likely to lose the higher-modulation frequencies, one should expect predictions of the BILD based on the MTF to be smaller than those based on the articulation index, and, therefore, smaller than those observed. Unfortunately, there are, at present, insufficient empirical data on the importance of different modulation frequencies in speech perception to make quantitative predictions of the BILD on this basis.

The fact that binaural sluggishness seems to be irrelevant to accurate predictions of the BILD consequently poses a problem for the idea that speech is heard via a temporally smeared representation of spectro-temporal modulations in interaural correlation. It is possible that this idea is flawed, since Yost (1985) has suggested an alternative hypothesis; binaural sluggishness might be related to the time it takes the binaural system to adapt to a changing masker. He pointed out that within the context of Durlach’s equalization-cancellation model (Durlach, 1972), such sluggishness could be represented as the time required for the system to calculate a new equalization operation. If binaural sluggishness took that form, it would be important for changing noises, but not for changing signals, like speech, because the equalization operation depends on the interaural configuration of the noise and that remains the same throughout the stimulus. In that case, the binaural spectrogram idea proposed here would be quite inappropriate. It should be noted that attempts to measure binaural temporal resolution using signal-detection tasks have all used temporally changing masking noises, rather than changing signals (Bell, 1972; Grantham and Wightman, 1979; Shackleton and Bowsher, 1989; Kollmeier and Gilkey, 1990; Culling and Summerfield, 1998). Thus, Yost’s account of binaural sluggishness does not affect the prediction of the BILD and is consistent both with Levitt and Rabiner’s findings, as well as with the binaural-temporal-resolution literature. However, Kohlrausch (1990) has found some evidence against Yost’s account. He showed that when the interaural phase of a masker is different in different frequency regions, listeners have no difficulty detecting antiphaseically presented tones in either region, when there is trial-by-trial uncertainty as to the tone frequency. He also noted that in a condition where So and S1 thresholds were measured at the same frequency, with trial-by-trial uncertainty as to the signal phase, the results could only be accounted for if the system was employing different equalization delays in parallel within the same frequency channel.

The present investigation compares two hypotheses: (1) binaural sluggishness does not extend to the discrimination of signals according to their spectro-temporal patterns; (2) the modulation frequencies in speech that make the most important contribution to intelligibility are too low to be affected by binaural sluggishness. Sections A–C of this introduction briefly outline the evidence that the binaural system is sluggish, the evidence that interaural decorrelation is interpreted by the binaural system as the presence of a masked signal, and the implications of combining these concepts into a representation of speech as recovered from masking noise by the binaural system—a “binaural spectrogram.” Section D outlines the experiments which follow.

A. Binaural sluggishness

The binaural system is unable to follow rapid changes in the interaural parameters of the stimulus. This sluggishness can be illustrated by asking listeners to discriminate between a stimulus whose binaural parameters are sinusoidally modulated in some way and a comparison stimulus which has the same average binaural characteristics, but unmodulated. If a binaural parameter is modulated at 10 Hz or more, the stimulus is much more difficult to discriminate from the unmodulated standard than at low modulation rates (Grantham and Wightman, 1978; Grantham, 1982). Recently, there have been several attempts to characterize binaural sluggishness as a moving-average filter, termed a binaural temporal window, which integrates information over time according to a weighting function that defines the window. These studies have sought to measure both the shape and equivalent rectangular angular duration (ERD) of the window (Kollmeier and Gilkey, 1990; Culling and Summerfield, 1998) or only its ERD (Akeroyd and Summerfield, 1999). The ERDs measured in these studies range from 40–200 ms. For the purposes of the present investigation, the measurements of Culling and Summerfield will be employed, since these authors measured both the shape and the ERD of the window, and since their method afforded greater protection against the effects of off-time listening (Moore et al., 1988; Plack and Moore, 1990) than that of Kollmeier and Gilkey. The window measured by Culling and Summerfield had an ERD of about 110 ms, and a Gaussian shape. The modulation transfer function for such a window, which is shown in Fig. 1, has a slope of about –18 dB/oct. It is worth noting for future reference that by 10 Hz (roughly the frequency at which the binaural beat is no longer heard as a moving stimulus) the attenuation is 28 dB.

B. Interaural decorrelation, the BMLD, and dichotic pitches

A number of studies have converged on the idea that signals in noise can be detected by the binaural system via
the differences in interaural correlation that they induce (Osman, 1971; Gabriel and Colburn, 1981; Durlach et al., 1986; Koehnke et al., 1986; Jain et al., 1991; Bernstein and Trahiotis, 1992, 1996a,b; Culling and Summerfield, 1995; Culling et al., 1998a,b,c). In particular, Jain et al. showed a correspondence between the amount of decorrelation that was induced by a pure-tone signal at masked threshold in a 1/10th-octave band of noise and the amount of decorrelation which could be detected in a just-noticeable-difference task, where correlation was controlled directly by mixing noises. Culling et al. (1998a,b) showed that this quantitative correspondence in the data is complemented by qualitative similarities in the perceptual experience. They showed that a group of phenomena known as dichotic pitches could be explained as illusions produced by the mechanism of binaural unmasking, since each of these phenomena involved stimuli which varied in correlation across frequency; although no ‘‘signal’’ is added to the noise, an additional sound is heard whose pitch corresponds to the interaurally decorrelated frequencies. Finally, Culling et al. (1998c) have found that the perceived loudness of decorrelated bands of noise fringed by correlated noise is consistent with the idea that decorrelation acted as a perceptual surrogate for signal intensity. Specifically, cumulative d’ for binary discriminations based on loudness was an approximately linear function of the ‘‘equivalent signal-to-noise ratio’’ in dB, where equivalent signal-to-noise ratio (SNR) was the SNR at which a signal in the NoS π configuration would produce the interaural decorrelation used in the experiment.

In the past, models of binaural unmasking have generally featured a process of applying internal delays to compensate for the interaural delays or phase shifts that have been applied to the noise, followed by some assessment of the correlation. If the noise is coherent, but a signal is present, the correlation will deviate from unity. For instance, in Durlach’s equalization–cancellation (E-C) model, delays (or phase shifts) are used to ‘‘equalize’’ the stimulus at the two ears and then the degree of correlation is assessed by a ‘‘cancellation’’ process that subtracts the stimulus at one ear from that at the other. In most cases, the greater the residue from cancellation, the greater the decorrelation. Osman (1971) developed a similar scheme in which the correlation after equalization was used directly as the decision variable. In order to account for BMLDs when the frequency and phase of a 20-ms signal were uncertain, and could not be predicted from the noise (which had different phases at each frequency), Kohlrausch (1990) suggested a modification to Durlach’s E-C scheme, in which all equalization operations are processed in parallel at all frequencies and equalization delays. Coming from a quite different perspective, Culling and Summerfield (1995) suggested that independently selected delays are employed in each frequency channel in order to account for listeners’ inability to perceptually segregate ‘‘whispered’’ vowels. Colburn (1973, 1977) employed a similar delay-only scheme, except that the within-channel cross product was calculated at a range of delays, and deviations of these cross products from reference values (supplied by the nonsignal interval, or by other frequency channels) indicated the presence of a signal. Such models may be most simply described as systems that are sensitive to within-channel incoherence in the stimulus. That is to say that the output at each frequency reflects the deviation from unity of the maximum in the cross-correlation function as a function of internal delay. When we refer to ‘‘decorrelation’’ we refer to such incoherence, and conversely when we refer to ‘‘correlation’’ we refer to coherence (i.e., the maximum value in the interaural cross-correlation function); the application of an optimum internal delay is assumed. For NoS π, the optimum delay is always zero. Given that reductions in correlation (from a coherent background) are widely thought to be the relevant cue for binaural detection, it seems sensible to calculate decorrelation directly, rather than use the various surrogate measures which have been used in previous models.

C. Binaural spectrograms

The idea of a binaural spectrogram is that the binaural system derives a spectro-temporal representation of interaural correlation as a function of time and frequency and that this representation provides a parallel input to the speech recognition system. In order to produce a perceptually valid binaural spectrogram, one must complete four stages of processing: (a) simulate realistic frequency selectivity; (b) derive the within-channel interaural decorrelation as a function of time; (c) smooth that function in accordance with the binaural temporal window; (d) transform the resulting function according to the sensitivity of the binaural system to decorrelation. Data have already been published which enable the first three stages to be implemented. The fourth stage has recently been addressed by Culling et al. (1998c). However, only the first three stages will be implemented here (Fig. 2).

The width of binaural critical bands appears to be similar to that of monaural critical bands (Kohlrausch, 1988; Kollmeier and Holube, 1992), so a conventional auditory filterbank such as that developed by Patterson et al. (1987, 1988) may be used as a front end [stage (a)]. The temporally smoothed within-channel correlation [stages (b) and (c)] may be implemented as follows. The interaural correlation is calculated directly from the stimulus (signal and noise) within a series of overlapping binaural temporal windows. The Gaussian window, derived by Culling and Summerfield (1998) at 500 Hz and at a noise spectrum level of 40 dB/SPL/Hz is employed throughout this paper. As published, the window represents the weighting of interaural correlation as a function of time and not the weighting of raw waveforms whose correlation is to be determined. In the process
of calculating correlation, the waveforms are multiplied together, multiplying, at the same time, any weighting that has been applied to them. For this reason, the weights of the window were square-rooted before being applied to both input waveforms, and the correlations between the corresponding windowed segments calculated. Finally, in order to gain a representation that reflects listeners’ sensitivity to interaural correlation, the temporally smoothed interaural correlation function derived from a series of overlapping windows may be scaled according to the relationship between decorrelation and sensitivity [stage (d)].

Figures 3 and 4 illustrate the effects of the monaural and binaural temporal windows [stages (a)–(c)] on the representation of speech. The speech segment contains the words ‘pirate’s gold.’ For Fig. 3 the speech segment is processed with no noise. A psychoacoustic spectrogram was generated using the method illustrated in Fig. 2. The speech segment was first added to speech-shaped noise in the NoS filter configuration at an overall signal-to-noise ratio of $-15$ dB. A binaural psychoacoustic spectrogram was generated using the method described above and illustrated in Fig. 2. Figure 4 differs dramatically in appearance from Fig. 3 because the binaural temporal window is an order of magnitude longer than the monaural temporal window and so greatly attenuates the representation of higher-modulation frequencies. Notice, in particular, that the formant movements which are clear in Fig. 3 at the onset of and during enunciation of the diphthong /aʊ/ of the word ‘gold’ are almost entirely smoothed out in Fig. 4. Thus, it appears that a prediction which comes from this representation is that the binaural system can recover information chiefly about the frequency and amplitude of the first formant in each syllable, but not about any formant movements.

It should be noted that the representation in Fig. 4 lacks any consideration of the relative effectiveness of the binaural system at different frequencies. Information above 1 kHz ($\approx 15$ ERBs) should be greatly attenuated. It is thought that the underlying reason for this lack of sensitivity is that the binaural system is relying upon the normalized interaural correlation (rather than the normalized interaural covariance) of the signal envelope at high frequencies (van de Par and Kohlrausch, 1995; Bernstein and Trahiotis, 1996a,b).

Formant movements are widely regarded as very important in speech perception. They provide information about both consonants and vowels. For instance, Strange et al. (1983) have shown that the formant movements entering and

FIG. 2. The stages of making a binaural spectrogram. The left- and right-hand channels are filtered by a gamma-tone filterbank (Patterson et al., 1987, 1988) and overlapping temporal windows are extracted. The windows of corresponding frequency and timing from the two ears are cross-correlated. The differences between the maximum interaural cross correlations and 1.0 as a function of delay for each frequency and time form the binaural spectrogram.
leaving a steady-state vowel (from a ‘‘briskly’’ spoken CVC syllable) are more useful to listeners in identifying the vowel than the steady-state portion. The binaural spectrogram suggests that such information would not be recovered from noise by the binaural system.

D. The present investigation

The first experiment verifies the binaural spectrogram concept by testing whether binaural sluggishness can be observed in the discrimination of signals according to their spectro-temporal patterns. The threshold levels for discriminating different repeated signal patterns (ascending vs descending arpeggios) were measured in the NoSo and NoS\(\pi\) configurations. Where a lower discrimination threshold is observed in NoS\(\pi\) than in NoSo, the binaural system may be presumed to be responsible. By varying the rate at which the spectro-temporal pattern of the signal is repeated, the modulation frequencies in the signal and, therefore, the rate of modulations in interaural correlation for the NoS\(\pi\) stimuli can be controlled. It will be shown that the difference between NoSo and NoS\(\pi\) discrimination thresholds diminishes with increasing repetition rate consistent with a loss of information about the modulations in interaural correlation furnished by the binaural system. At high repetition rates the binaural system is too sluggish to substantially assist the listener, resulting in similar discrimination thresholds for NoSo and NoS\(\pi\).

Having established that changing signals do display effects of binaural sluggishness, the second experiment explores the possibility that, for speech, the most important modulation frequencies are sufficiently low that they are not affected by binaural sluggishness. This experiment tests whether speech perception in noise will begin to show the detrimental effects of binaural sluggishness if the modulation frequencies in speech are artificially increased.

I. EXPERIMENT 1

A. Stimuli

Experiment 1 tests the hypothesis that binaural sluggishness will be observed when listeners try to distinguish two masked signals which have different spectro-temporal patterns. For this purpose, two simple spectro-temporal patterns were used. These signals were three-tone arpeggios, made up from pure-tone bursts of variable duration (32–100 ms) which included 5-ms, raised-cosine onset/offset ramps. Each tone began at the offset of the previous tone, so that a shorter tone duration would result in a more rapid arpeggio. The arpeggios began on a randomly selected tone and were repeated throughout a 1.6-s, white noise burst (0–10 kHz). The three pure tones had frequencies of 400, 500, and 625 Hz. The onsets and offsets of the resulting 1.6-s sequence of arpeggios was ramped on and off using a 50-ms, raised-cosine function, while the noise was ramped on and off more rapidly, using a 10-ms raised cosine function. The random starting tone and the slower ramping of the signal were intended to remove listeners’ opportunity to exploit the initial or final tone in the sequence as a cue to the direction of frequency change. The durations of the constituent tones (held constant during a given threshold run) were 32, 40, 50, 64, 80, or 100 ms, resulting in arpeggios of three times these durations. The corresponding rates of repetition were, therefore, 10.41, 8.33, 6.67, 5.21, 4.17, and 3.33 arp/s. The signals were presented in either the NoSo or NoS\(\pi\) binaural configurations. The noise was presented at a spectrum level of 38 dB(SPL)/Hz.

As well as measuring the masked thresholds for 70.7% discrimination performance for different tone durations, the masked detection thresholds were also measured. This test was run in exactly the same way, except that the downward arpeggios were removed, so that listeners discriminated masked upward arpeggios from noise alone. Measurements of the detection thresholds facilitated the calculation of discrimination thresholds expressed in sensation level.2

All stimuli were generated on-line using a TDT AP2 array-processor card and presented to listeners in a sound-treated chamber via a TDT System-2 psychoacoustic rig (DD1, PA4, FT5-9, HB6) and Sennheiser HD414 head-phones.

B. Procedure

Listeners attended a number of 1-h practice sessions followed by eight experimental sessions. The first four experimental sessions measured discrimination thresholds and the second four measured detection thresholds. In each session, they completed 12 runs which covered all combinations of the six rates of arpeggios and the two binaural configurations. Each run was a 2-down/1-up adaptive threshold measurement (Levitt, 1971), with 14 reversals. The average signal level at the last ten reversals was taken as the threshold. The stepsizes were 4 dB until two reversals had been completed, and 2 dB thereafter. Each trial was a two interval, forced-choice (2I-FC) task, in which the listener was required to indicate the interval which contained ascending arpeggios. Trial-by-trial feedback was given. The listeners were trained on a predictable sequence of runs that had progressively faster arpeggios throughout the session, but were tested using a random sequence of runs. After collecting four thresholds for each condition in the discrimination task, four detection thresholds were measured for the same conditions. For this purpose the downward arpeggios were removed from each trial, so that listeners simply indicated the noise burst which contained an upward arpeggio. Four discrimination thresholds in dB sensation level (SL) were derived by subtracting the discrimination thresholds from the corresponding detection thresholds.3

The discrimination task was found to be very difficult by most listeners. Often the adaptive staircase returned to the initial (and maximum) signal level of 84 dB SPL during a run at one of the fastest presentation rates. Three listeners could not discriminate the direction of movement accurately at the lowest presentation rate and at high signal-to-noise ratios, and were immediately rejected. Another four were trained for many hours without achieving stable thresholds at the highest rates. Five listeners eventually achieved stable thresholds across all presentation rates. Of these, only two (AM and MT) were able to do this without at least 5–10 h of
training. Four listeners produced a complete data set with stable thresholds throughout. Their data are presented below.

C. Results

Mean thresholds, averaged across four runs, for the four listeners are plotted in separate panels of Figs. 5, 6, and 7. Figure 5 shows the masked discrimination thresholds in dB. For each listener, the discrimination thresholds for NoSo and NoS\textsubscript{p} differ by 12–14 dB at a rate of 3.33 arps/s, but this binaural advantage reduces to 3–5 dB at 10.4 arps/s. The difference in discrimination level at 3.33 arps/s is therefore similar to the BMLD for detection of static tones in noise, but much smaller when the repetition rate is higher. Although both NoSo and NoS\textsubscript{p} thresholds increase with presentation rate, NoS\textsubscript{p} thresholds increase more. An analysis of variance (ANOVA) covering the four runs, six presentation rates, and two binaural conditions showed significant main effects of rate of presentation \( F(5,15)=5.42, p<0.0001 \) and of binaural condition \( F(1,3)=466, p<0.0005 \), and a significant interaction between those two \( F(5,15)=117, p<0.0001 \). There were no effects associated with the ‘run’ factor. Tukey HSD pair-wise comparisons showed that the difference between NoSo and NoS\textsubscript{p} was significant at the 5% level or less at all presentation rates.

Figure 6 shows the detection thresholds for ascending arpeggios using the same scale. The detection thresholds also increased as a function of presentation rate; the most rapid sequences gave detection thresholds which were up to 7 dB higher than the slowest for both NoSo and NoS\textsubscript{p}. The slope of this increase in threshold with presentation rate varies across the listeners, but is similar for each listener in NoSo and NoS\textsubscript{p}.

Figure 7 shows the discrimination thresholds in dB SL \( \text{SL} \) (i.e., the thresholds from Fig. 6 have been subtracted from those from Fig. 5). Before the detection thresholds were subtracted, the discrimination thresholds in the NoSo condition showed a substantial increase with presentation rate. Since detection thresholds increase with presentation rate, discrimination thresholds in NoSo increase less when measured in sensation level than when measured in absolute level. The thresholds in sensation level are similar for NoSo and NoS\textsubscript{p} at low presentation rates and separate as the presentation rate increases for all listeners. Listeners can make the discrimination at a sensation level of 5–10 dB in the NoSo condition across all presentation rates. In the NoS\textsubscript{p} condition, however, the threshold is similar to NoSo at the lowest presentation rate but is substantially higher at higher presentation rates; the listeners can discriminate the direction of the fast-
D. Discussion

The results show that the binaural advantage for discriminating two simple spectro-temporal patterns in noise can be greatly reduced if the patterns change rapidly. This reduction is consistent with a lack of temporal resolution for binaural processing. Binaural sluggishness has therefore been observed in the discrimination of spectro-temporal patterns among masked signals. This result is inconsistent with Yost’s (1985) suggestion that binaural sluggishness might reflect the time taken for the binaural system adapt to a changing masker, since the masker was constant in this experiment. The data therefore corroborate Kohlrausch’s (1990) conclusions on this matter, using a quite different approach. However, while the reduction in binaural advantage was progressive, the binaural advantage was not completely abolished at the highest rate. This aspect of the results appears inconsistent in some respects with the Gaussian window shape of 110-ms ERD measured by Culling and Summerfield (1998). Such a window should result in a high degree of attenuation (about 28 dB) for modulations of correlation at 10.4 Hz (cf. Fig. 1). The point may perhaps be better illustrated using the binaural spectrogram itself. Figures 8, 9, and 10 show binaural spectrograms of descending arpeggios with presentation rates of 3.3, 5.2, and 10.4 arp/s, respectively. These spectrograms have been plotted as grayscale plots with discrete contours so that the exact correlations are shown explicitly at the boundaries. The signal-to-noise ratio used is indicated by arrows in Fig. 5. At this signal-to-noise ratio, the presence of the signal causes substantial decorrelation of the masker for all rates, consistent with detection performance well above threshold. However, the representation of the spectro-temporal pattern of decorrelation depends upon the arpeggio rate; the binaural spectrogram contains clear regular modulations for downward sweeps at 3.3 arp/s, rather less distinct modulation for downward sweeps at 5.2 arp/s, and no discernible detail at 10.4 arp/s. Figure 10 shows some random fluctuations in correlation. These are caused by fluctuations in the level of the noise; whenever the noise fluctuates downward in level, the decorrelation caused by the presence of the signal increases. According to a visual inspection of Fig. 10, the binaural spectrogram model predicts that listeners should be unable to use binaural cues at 10.4 arp/s and so should have similar thresholds for NoSo and NoSπ conditions at that presentation rate. These thresholds in fact differ by 3–5 dB, suggesting that the model overpredicts the increase in thresholds with increasing presentation rate.

The fact that the model overpredicts this increase was confirmed by some more formal modeling. Stimuli were created with signals at the listeners’ average discrimination threshold from the experiment for presentation rates of 3.3 and 10.4 arp/s. The interaural correlation at the output of a pair of 500-Hz gamma-tone filters was measured during the 500-Hz tone bursts. This average correlation value was then used to create a boxcar correlation modulation function with a 2/3 duty cycle (i.e., with a correlation of 1.0 for 2/3 of a cycle). This function represented the idealized modulation in interaural correlation at the output of a matched pair of auditory filters with 500-Hz center frequencies when listening to the arpeggios. The function was convolved with the binaural temporal window and the modulation of the correlation at the output of the window assessed. At 3.3 arp/s the correlation varied between 0.998 and 0.941, a difference of 0.057. This value is comparable with just-noticeable differences (jnd’s) reported in the literature for a reference correlation close to 1.0 (Pollack and Trittipoe, 1959a; Gabriel and Colburn, 1981; Koehnke et al., 1986; Culling et al., 1998c), which vary between 0.03 and 0.005. At 10.4 arp/s the correlation varied between 0.628 and 0.572, a difference of 0.055. The range of correlations at the outputs of the temporal window is, therefore, very similar at each presentation rate. However, it is well established that listeners are progressively less sensitive to changes in interaural correlation as the reference correlation is reduced from 1.0 (Pollack and Trittipoe, 1959a; Culling et al., 1998c). So, at threshold, there should be a larger range of correlation change at the output of the temporal window for the faster arpeggios than for the slower ones due to their lower average interaural correlation (~0.6). Since this is not observed, the listeners are coping better with the increase in presentation rate than the model would predict. This said, the model did predict a reduction in discrimination level difference over the approxi-
mate range of presentation rates over which this reduction took place.

The increase in detection thresholds with rate of presentation (Fig. 6) was an unexpected result. The effect remains when the reductions in signal level produced by the onset and offset ramps is accounted for. Some informal follow-up experiments (using listener JC, the first author) suggest that the effect is produced by the use of multifrequency spectro-temporal sound patterns. When two of the tones were switched off, there was no increase in threshold for the remaining tone as its repetition rate was increased. It seems likely that the effect arises from listeners attempting to attend to the tone pattern one tone at a time; if listeners can succeed in attending to the appropriate frequency channel throughout part of the stimulus, then they may hear it better. Of course, the listener must have heard the preceding tones in order to know what frequency will come next, but since the stimulus is extended in time, listeners may have the opportunity to confirm or refute a hypothesis about how the frequency is changing during the course of the stimulus, by attempting to follow the frequency changes. When the sequence is too rapid for listeners to switch their attention between frequency channels, the strategy is confounded and thresholds increase. The effect may be worthy of further study.

The failure of some listeners to perform the task at all also deserves some comment. Two listeners could not discriminate the upward from downward arpeggios not only at the slowest rates presented in the experiment, but even at 2.08, 1.67, and 1.04 arp/s specially prepared by the experimenter for initial training. Lower rates than this were impractical without increasing the stimulus duration, as fewer than two arpeggios were already being presented at 1.04 arp/s. In this situation, it is very difficult to know whether listeners have simply not grasped the instructions, although one would expect the trial-by-trial feedback to resolve any confusions provided that the stimuli were discriminable. It seems that some listeners with apparently normal hearing and speech can find abstract spectro-temporal sound patterns very difficult to discriminate.

II. EXPERIMENT 2

Experiment 2 investigated the effect of speech rate on speech-reception thresholds (SRTs) in NoSo and NoS\(\pi\) binaural configurations. If the binaural system responds sluggishly to the changes in interaural correlation induced by the fluctuating presence of speech energy in the NoS\(\pi\) configuration, and if detection of those correlation changes is the basis of the binaural intelligibility level difference (BILD), then high articulation rates should reduce the BILD.

A. Stimuli

The speech consisted of recorded sentences from the Harvard sentence list (Rothaeuser et al., 1969). The speech was processed digitally in order to increase the articulation rate without altering the pitch, using a method similar to the phase vocoder. This acceleration of the articulation rate was implemented using the Hilbert transform. The procedure was as follows. (1) The digital waveforms were filtered by the Patterson et al. (1987, 1988) gamma-tone filterbank. (2) Each filtered channel was converted to an instantaneous amplitude and an instantaneous phase function. (3) The instantaneous phase function was unwarped to produce a monotonically increasing function. (4) The phase function was scaled down (by the reciprocal of the acceleration factor), in order to prevent an increase in carrier frequency. (5) The phase and amplitude functions were resampled at a reduced sampling rate in order to compress the same changes in frequency and amplitude over time into a smaller number of samples. (6) An inverse Hilbert transform recovered the accelerated waveforms. (7) All channels were summed back together to get the accelerated speech signal.\(^4\)

The masking noise was speech-shaped according to the long-term spectra of the speech materials used in the experiment. Fifty samples of such noise were prepared. Speech and noise were prepared off-line using WAVE software (Culling, 1996) and mixed adaptively during the experiment using a TDT AP2 array processor.

B. Procedure

The noise was mixed with the speech and presented to the listener at a signal-to-noise ratio determined by a procedure adapted from one developed by Zurek (1996). SRTs were measured in a series of 1-up/1-down adaptive runs to determine the 50%-intelligibility level (Levitt, 1971). Each run began with a sentence at a poor SNR. For this sentence, the listener was instructed to increase the signal level progressively and rehear the sentence until it was perceived as partially intelligible, and then to attempt a transcription. By this means, the adaptive staircase was started at approximately the 50%-intelligibility level without wasting any speech material. The listener entered the transcript on a computer terminal. When the transcript was complete, the listener pressed “return” and the computer gave the listener the actual transcript of the speech with the keywords in upper case. The listener self-marked the number of correct keywords out of the five in each sentence and the computer adjusted the signal level by 2 dB for the next trial. For subsequent trials, the listener heard each sentence only once and immediately offered a transcript. The SNR was increased for two or fewer keywords correct and decreased for three or more correct. The listeners were not informed of this criterion, but were aware that the computer was using their self-reported performance to control the speech level and also that the entire transaction was both visible on the experimenter’s monitor and being logged. A complete adaptive run consisted of a total of ten sentences. The new signal level selected after each trial was averaged over the last eight trials to give the measured speech-reception threshold for that run.

The listeners attended either six or ten 1-h sessions. All listeners began with two sessions of practice with the experimental procedure for unprocessed sentences in the NoSo and NoS\(\pi\) configurations. This practice familiarized listeners with the procedure, ensuring that their judgments of partial intelligibility at the start of each run were acceptably accurate, and also gave an initial estimate of the binaural intelligibility level difference (BILD). At this stage, one listener showed no BILD and was rejected from the study. The other listeners averaged a BILD of 5.4 dB. Six listeners went
straight on to the experimental sessions, and six listeners were given further training designed to acclimatize them to accelerated speech. This training consisted of listening to a series of story tapes for a recently published thriller, using a variable-speed tape player (Radio Shack VSC-2002). This machine was also equipped with a pitch control (Variable Speech Control Co.), which could be used to offset the pitch changes produced by changing the tape speed. These listeners were instructed to attend to the story, which none of them had heard before, and to increase the tape speed progressively until they were following the story at double the original speed. All six of these listeners increased the speed to double by the end of the first acclimatization session and listened comfortably at double speed throughout the following three sessions.

In the four experimental sessions, listeners completed a total of 44 runs. The first two runs in the first session were further practice, but using accelerated speech material. The remaining 42 runs were seven repeated measures for the six conditions (3 acceleration factors × 2 binaural conditions). The speech materials were new to the listeners, and were counterbalanced across the six conditions, so that each sentence was presented in each of the six conditions; for a given set of six listeners, each sentence was presented once in each of the six conditions, each time to a different listener.

C. Results

Figure 11 shows the results of experiment 2 for the two groups of listeners combined. The results are averaged across all 12 listeners because they would otherwise be distorted by differences in the intrinsic intelligibility of the different sentence lists (only across a group of six listeners are all lists used for all six conditions) and because an analysis of variance, covering the seven repeated-threshold measurements, the three speech rates (1, 1.5, and 2 times original speed), the two binaural conditions (NoSo/NoSπ), and the two training regimes (with/without accelerated-speech acclimatization) showed no effect of acclimatization to accelerated speech.

Listeners produced SRTs that were 3–5 dB lower for NoSπ (filled symbols) than for NoSo (open symbols) [F(1,10)=27, p<0.0005]. An analysis of simple main effects shows that the difference was significant at each speech rate [at original speed, F(1)=16.8, p<0.002; at 1.5 times original speed, F(1)=9.4, p<0.02; at twice original speed, F(1)=5.0, p<0.05]. Meanwhile, their thresholds in both conditions rose with increasing articulation rate by around 8 dB [F(2,20)=172, p<0.0001]. The increase in SRT was around 2 dB greater for NoSπ than for NoSo, as reflected by an interaction between acceleration factor and binaural condition in the ANOVA [F(6,60)=2.6, p<0.05].

D. Discussion

In both experiments 1 and 2, the principal effect of interest was the interaction between rate of change in the signal and binaural condition. In experiment 1, increasing the repetition rate of the arpeggios substantially reduced the effect of the binaural configuration, broadly in agreement with the predictions of a binaural spectrogram model. However, in experiment 2, the most striking effect of accelerated speech was a 6–8-dB increase in SRTs. In addition to this effect there was a reduction in the binaural advantage for accelerated speech. However, the 6–8-dB increase in sensation level makes it difficult to interpret the reduction in binaural advantage with certainty. Thus, it would appear that while accelerated speech is difficult to understand in noise, this effect has relatively little to do with binaural sluggishness. This result was somewhat unexpected, and requires explanation.

Given that experiment 1 is persuasive in its demonstration of an effect of binaural sluggishness on spectrottemporal pattern discrimination, the most likely explanation for the relative indifference of the binaural system to increased speech rate is that the binaural system is still able to recover useful information which is within the temporal window’s MTF. In order to assess the likelihood of this explanation, one needs to know what modulation frequencies are present in speech and how important they are to intelligibility. Plomp (1983) measured the average modulation spectrum for ten male voices. He found that the modulation spectrum was similar in all measured frequency bands; in each 1/3rd-octave band the modulation spectrum peaked at about 4 Hz (approximately the syllable rate), but featured extensive shallow slopes towards higher and lower modulation frequencies. A similar demonstration appears in Houtgast and Steeneken (1985). The dominant modulation frequencies of speech are therefore quite low compared to the passband of the temporal window, and doubling the articulation rate would not remove them. However, these frequencies might not be the most important.

Drullman et al. (1994a,b) have investigated the role of different modulation frequencies in speech intelligibility. In their first paper (1994a), they temporally smeared speech using each of a wide range of low-pass cutoff frequencies (4, 8, 16, 32, and 64 Hz) before adding noise. They found that the 16-, 32-, and 64-Hz cutoffs produced very similar SRTs. The 8-Hz cutoff increased the SRT by about 1 dB and the 4-Hz cutoff increased it by a further 4 dB. Their experiment suggested, therefore, that frequencies between 4 and 8 Hz are
quite important to the intelligibility of speech in noise, although it was not clear whether they are more or less important than still lower frequencies. Their subsequent paper (1994b) used high-pass filtering of the modulation spectrum in order to address this question. By combining the results of the two studies, they demonstrated that the most important modulation frequency was, in fact, about 8 Hz.

Their first experiment was similar, in a sense, to experiment 2 of the current study: Drullman et al. directly smeared the speech, while experiment 2 assumed that such smearing would be a side effect of binaural processing; Drullman et al. varied the low-pass cutoff of the modulation spectrum, while Experiment 2 varied the speech rate. In either case, speech was heard in noise with a low-pass-filtered modulation spectrum. If we say that the binaural system recovers no information about the modulation spectrum above about 10 Hz (28-dB attenuation according to Fig. 1), the information in the original modulation spectrum is effectively low-pass filtered at 10, 6.7, and 5 Hz for acceleration factors of 1, 1.5, and 2, respectively. Experiment 2 found only partial abolition of the BILD for an equivalent 5-Hz low-pass cutoff. Assuming that the original speech rate of our materials was similar to Drullman’s, it is surprising that no more substantial effect was observed in experiment 2. It may be that most of the modulation frequencies above 4 Hz that are important to intelligibility are in higher spectral regions than those recovered by the binaural system. In any case, the present results indicate that binaural sluggishness does not limit the ability of the binaural system to facilitate speech understanding in noise at normal speech rates.

It is possible that across-frequency processing may help explain the discrepancy between the results obtained with tone patterns and with speech. Gordon (1997a,b) has shown that above-threshold information from remote spectral regions can “protect” temporally coherent signals from masking. He observed a modest reduction in threshold (~2 dB) for a brief signal in low-pass noise when a cosignal was simultaneously presented at higher frequencies. The cosignal was presented in both the signal and nonsignal intervals and apparently served only as a temporal marker for the signal, since discrepancies between the onset and offset times of the signal and cosignal destroyed the effect. Conceivably, a similar effect may occur when listening to speech in noise. The low-frequency, low-modulation rate information recovered by the binaural system may act as a cosignal which assists detection of temporally coherent parts of the speech signal in higher spectral regions. Akeroyd et al. (1998) have recently demonstrated that the auditory system can integrate speech information furnished simultaneously by monaural and binaural cues and the requirement for temporal coherence should be fulfilled by speech signals because all frequencies tend to modulate coherently at 4 Hz or so due to the enunciation of syllables.

The experiments and modeling described here demonstrate that binaural sluggishness can affect the auditory system’s ability to extract fine temporal detail from signals in noise. However, the indications are that this effect is of marginal importance to the BILD. It is not yet clear why the effect of binaural sluggishness does not have a more detrimental effect on speech understanding. A better understanding of how binaural information is employed in the unmasking of speech will be required to resolve this issue.

III. CONCLUSIONS

1. Binaural sluggishness reduces the binaural advantage for discriminating spectro-temporal patterns in noise when those patterns change rapidly.

2. Accelerated speech is more difficult to understand in noise than the same speech at the original rate regardless of binaural configuration.

3. The speech signal may contain information that is sufficient to facilitate understanding at low-modulation frequencies in the 200–1500-Hz spectral region where the binaural system is effective.

ACKNOWLEDGMENTS

We would like to thank Gerald Kidd and Christine Mason for many helpful suggestions about the design of the experiments, Katrin Krumbholz for advice on the Hilbert transform, and Armin Kohlrausch and an anonymous reviewer for their thorough and thought-provoking reviews. Work supported by an M.R.C. Career Development Award (J.F.C.) and by NIDCD (H.S.C.).

1. In each of these studies stimulus artifacts enabled listeners to make discriminations at modulation rates much higher than 10 Hz. Generally performance reached a trough in the 10–50 Hz region.

2. Thresholds in sensitivity level were calculated because Yost (1997) has pointed out that binaural advantages can become smaller as sensation level is increased. Since there was a general trend for thresholds to increase for faster presentation rates, it was thought possible that the increase in sensation level might itself reduce the binaural advantage, rather than binaural sluggishness per se. In fact, we found that the detection thresholds also increased with presentation rates, so the increase in sensation level was small.

3. Listener MB was only available for two sessions of detection threshold measurements, so the two resulting thresholds for each condition were subtracted from two different discrimination thresholds.

4. The resulting accelerated speech was of very high quality. Occasionally, the excitation had a slightly gurgling quality, which, upon investigation, could be attributed to the broad high-frequency channels of the gammatone filterbank. Because these channels admitted a number of frequency components, the instantaneous amplitude function displayed modulation at the original fundamental frequency. After resampling, this modulation was increased in frequency. Thus, in the finished stimuli the fundamental frequency information at low frequencies was fully controlled by scaling the phase of the isolated components in these channels, and was identical to the original, while the high frequencies were giving amplitude modulation cues to a higher fundamental. The perceptual salience of this conflicting information was very low due to the dominance of low frequencies in pitch perception.

5. The real-time pitch control of this device appeared to be based on the pitch period deletion principle rather than a phase vocoder. Although the pitch was controlled by a slider, the pitch changes were discrete.


