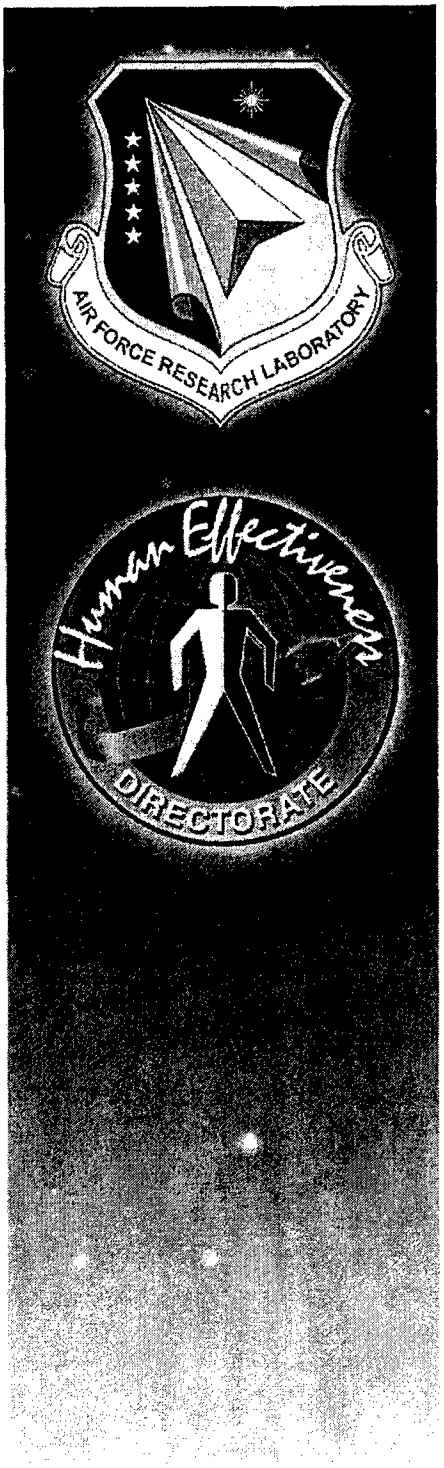
AFRL-HE-WP-JA-2005-0012



AIR FORCE RESEARCH LABORATORY

Across-ear Interference from Parametrically Degraded Synthetic Speech Signals in a Dichotic Cocktail-party Listening Task

Douglas S Brungart
Brian D. Simpson

Air Force Research Laboratory

Christopher J. Darwin

University of Sussex
Falmer, BN1 9QH, England

Tanya L. Arbogast
Gerald Kidd, Jr.

Hearing Research Center
Boston University
635 Commonwealth Avenue
Boston MA 02215

January 2005

Interim Report for October 2004 to January 2005

Human Effectiveness Directorate
Warfighter Interface Division
Battlespace Acoustics Branch
2610 Seventh Street
Wright-Patterson AFB OH 45433-7901

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Air Force Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-JA-2005-0012

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

//Signed//

BRADFORD P. KENNEY, Lt Col, USAF
Deputy Chief, Warfighter Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
January 2005	Interim	October 2004 to January 2005

4. TITLE AND SUBTITLE
Across-ear Interference from Parametrically Degraded Synthetic Speech Signals in a Dichotic Cocktail-party Listening Task

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Douglas S. Brungart, Brian D. Simpson (AFRL)
Christopher J. Darwin (Univ. of Sussex)
Tanya L Arbogast, Gerald Kidd Jr. (Boston Univ.)

5d. PROJECT NUMBER
2313

5e. TASK NUMBER
HC

5f. WORK UNIT NUMBER
52

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory, Human Effectiveness Directorate
Warfighter Interface Division
Battlespace Acoustics Branch
Air Force Materiel Command
Wright-Patterson AFB OH 45433-7901

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL-HE-WP-JA-2005-0012

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Originally published in The Journal of the Acoustical Society of America

14. ABSTRACT Recent results have shown that listeners attending to the quieter of two speech signals in one ear (the target ear) are highly susceptible to interference from normal or time-reversed speech signals presented in the unattended ear. However, speech-shaped noise signals have little impact on the segregation of speech in the opposite ear. This suggests that there is a fundamental difference between the across-ear interference effects of speech and nonspeech signals. In this experiment, the intelligibility and contralateral-ear masking characteristics of three synthetic speech signals with parametrically adjustable speech-like properties were examined: (1) a modulated noise-band (MNB) speech signal composed of fixed-frequency bands of envelope-modulated noise; (2) a modulated sine-band (MSB) speech signal composed of fixed-frequency amplitude-modulated sine waves; and (3) a "sinewave speech" signal composed of sine waves tracking the first four formants of speech. In all three cases, a systematic decrease in performance in the two-talker target-ear listening task was found as the number of bands in the contralateral speech-like masker increased.

15. SUBJECT TERMS
Speech perception, multitalker listening, cocktail-party effect

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Douglas S. Brungart
U	U	U	SAR	9	19b. TELEPHONE NUMBER (include area code) 937.255.3660

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18

This page intentionally left blank.

Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task

Douglas S. Brungart[a] and Brian D. Simpson
Air Force Research Laboratory, AFRL/HECB, 2610 Seventh Street, WPAFB, Ohio 45433

Christopher J. Darwin
University of Sussex, Falmer, BN1 9QH, England

Tanya L. Arbogast and Gerald Kidd, Jr.
Hearing Research Center, Boston University, 635 Commonwealth Avenue, Boston, Massachusetts 02215

(Received 5 March 2004; revised 1 November 2004; accepted 2 November 2004)

Recent results have shown that listeners attending to the quieter of two speech signals in one ear (the target ear) are highly susceptible to interference from normal or time-reversed speech signals presented in the unattended ear. However, speech-shaped noise signals have little impact on the segregation of speech in the opposite ear. This suggests that there is a fundamental difference between the across-ear interference effects of speech and nonspeech signals. In this experiment, the intelligibility and contralateral-ear masking characteristics of three synthetic speech signals with parametrically adjustable speech-like properties were examined: (1) a modulated noise-band (MNB) speech signal composed of fixed-frequency bands of envelope-modulated noise; (2) a modulated sine-band (MSB) speech signal composed of fixed-frequency amplitude-modulated sinewaves; and (3) a "sinewave speech" signal composed of sine waves tracking the first four formants of speech. In all three cases, a systematic decrease in performance in the two-talker target-ear listening task was found as the number of bands in the contralateral speech-like masker increased. These results suggest that speech-like fluctuations in the spectral envelope of a signal play an important role in determining the amount of across-ear interference that a signal will produce in a dichotic cocktail-party listening task. [DOI: 10.1121/1.1835509]

PACS numbers: 43.66.Pn, 43.66.Rq, 43.71.Gv [AK] Pages: 292–304

I. INTRODUCTION

Of all the difficult acoustic environments that occur in the everyday lives of human listeners, some of the most challenging involve the so-called "cocktail party problem" of listening to what one talker is saying when other talkers are speaking at the same time (Cherry, 1953). From a signal processing standpoint, this problem is extremely difficult, and even after years of intensive research the designers of automatic speech recognition systems still have not developed adequately robust algorithms for segregating speech in a wide variety of multitalker environments (Stern, 1998). Yet, normal-hearing human listeners are generally quite capable of understanding speech even in extremely complex situations that involve multiple simultaneous talkers in a reverberant environment.

Over the past 50 years, a great deal of research has been devoted to determining how listeners are able to achieve this success [see Yost (1997), Bronkhorst (2000), and Ebata (2003) for recent reviews of this literature]. In part, the answer lies in the inherent ability of human listeners to exploit differences in the voice characteristics of the different talkers, either in terms of fundamental frequency and intonation (Brokx and Nooteboom, 1982; Darwin and Hukin, 2000; de Cheveigne, 1993), vocal tract length (Darwin *et al.*, 2003), or overall speaking level (Egan *et al.*, 1954; Brungart, 2001b). In most situations, however, these monaural speech segregation cues are augmented by the binaural interaural level differences (ILDs) and interaural phase differences (IPDs) that occur when the target and interfering speech signals originate from different spatial locations relative to the listener (Bronkhorst and Plomp, 1988). These binaural difference cues enhance multitalker speech segregation in two ways: first, they introduce acoustic differences in the signals at the two ears that can be equivalent to as much as a 6–10-dB increase in the effective signal-to-noise ratio (SNR) of the target speech [e.g., see Zurek (1993)]; and second, they cause the target and masking signals to appear to originate from different locations in space, thus making it easier to selectively attend to one of the two speech signals (Freyman *et al.*, 1999).

In real-world listening environments, it is difficult to determine relative contributions these two types of binaural segregation cues make to the spatial unmasking of speech. Because all sound sources in realistic environments transmit some energy to each of the listener's two ears, some portion of the target speech signal will always be acoustically masked out by the interfering speech no matter how far apart the two sources are located. Thus, to the extent that listeners are unable to segregate widely separated speech signals in the free field, we cannot be sure whether the reason is because some portion of the target signal was obscured by the masker or because the two talkers did not "sound" far enough apart for the listener to perfectly segregate them.

[a]Electronic mail: douglas.brungart@wpafb.af.mil

There is, however, a somewhat artificial experimental manipulation that can be used to by-pass this inherent problem in real-world speech segregation. By presenting the target and masking signals "dichotically" over headphones (i.e., with one talker in one ear and one talker in the other ear), it is possible to generate a stimulus with two talkers who appear to originate from different places but have no acoustic overlap that could lead to energetic masking of the target.

Most of the experiments that have been conducted in these kinds of dichotic listening situations have shown that audio signals presented in one ear have little or no impact on the ability of normal-hearing adults to selectively attend to unrelated audio signals in the other ear. For example, Cherry (1953) has shown that a listener's ability to attend to a monaurally presented speech signal is unaffected by the presence of a distracting speech signal in the opposite ear. Other researchers have found similar results for the perception of dichotically separated speech signals (Drullman and Bronkhorst, 2000) and for the detection of tones in the presence of contralaterally presented random-frequency informational maskers (Neff, 1995; Wightman *et al.*, 2003). However, recent results have shown that the ability to ignore a distracting sound in the unattended ear can break down when a second distracting sound is also present in the same ear as the target signal. For example, Kidd and his colleagues (Kidd *et al.*, 2003) have shown that the presence of a random-frequency masker in the listener's unattended ear can sometimes impair the detection of a monaurally presented tone in the opposite ear when a second random-frequency masker is simultaneously presented in the same ear as the target tone. Similarly, Brungart and Simpson (2002) have shown that the presence of an interfering speech signal in the unattended ear can substantially impair the comprehension of a target speech signal in the opposite ear when a second independent interfering signal is simultaneously presented in the same ear as the target speech. Although other studies of dichotic speech perception have shown that listeners who are instructed to attend to a monaurally presented speech signal can be distracted by speech signals in the unattended ear that contain information that is surprising, unexpected, and/or relevant to the listener [such as an unexpected occurrence of the listener's first name (Moray, 1959; Wood and Cowan, 1995; Conway *et al.*, 2001)] or related in some way to the speech signal in the target ear [such as a midsentence swap between the signals in the target and unattended ears (Treisman, 1960)], historically there has been little evidence that irrelevant speech signals generate substantial amounts of across-ear interference in dichotic speech perception. The significance of Brungart and Simpson's (2002) finding is that it indicates that listeners in a dichotic listening task can be distracted by speech signals presented in the unattended ear even when those signals are unrelated to the target speech signals and completely devoid of any information that might be of interest to the listener outside the scope of the experimental task.

One intriguing aspect of Brungart and Simpson's dichotic speech segregation experiment was that significant across-ear interference occurred only for contralateral signals that were qualitatively "speech-like:" single-talker speech, multiple-talker speech, and time-reversed speech all caused across-ear interference, but speech-shaped noise did not. Furthermore, when the signal-to-noise ratio (SNR) in the target ear was less than 0 dB, time-reversed speech actually caused just as much across-ear interference as normal speech. Thus, it appears that, despite their obvious dissimilarities, normal speech signals and time-reversed speech signals share a common set of acoustic features that (a) interfere in some way with central speech processing, and (b) are not present in Gaussian noise. This conclusion suggests that some important insights into the processes that listeners use to segregate competing speech signals could be obtained by identifying the acoustic characteristics that cause audio signals to produce across-ear interference in dichotic listening. Furthermore, there is reason to believe that the underlying mechanisms that cause across-ear interference to occur for contralateral speech maskers in Brungart and Simpson's dichotic task might also extend into more realistic binaural listening situations where the target and masking signals are presented in different directions relative to the listener rather than in completely different ears. Indeed, such an effect might explain the relatively larger degradations in performance that have been shown to occur when a second speech masker is added to a stimulus containing two spatially separated competing speech sigals opposed to when a second noise masker is added to a stimulus containing a speech signal masked by a spatially separated noise source. Peissig and Kollmeier (1997), for example, found a 6.2-dB increase in speech reception threshold (SRT) when a second interfering talker was added to a speech signal masked by one competing talker, but only a 2-dB increase in SRT when a second interfering noise was added to a speech signal masked by one competing noise source. In a similar study, Hawley *et al.* (2004) reported a 9-dB increase in SRT with the addition of a second speech competitor to a stimulus containing two spatially separated speech signals, but only a 4-dB increase with the addition of a second noise competitor to a stimulus containing a target speech signal masked by a single spatially separated noise. Relatively large degradations in performance have also been shown to occur when a second interfering talker is added to a monaural stimulus containing two competing talkers (Brungart *et al.*, 2001; Hawley *et al.*, 2004). All of these results might be closely related to the Brungart and Simpson finding that listeners are able to use spatial location to segregate a target speech signal from one competing talker, but that they are unable to use location to segregate a speech signal from two competing talkers at different locations at the same time.

In this paper, we attempt to further explore the acoustic characteristics that cause a signal to interfere with dichotic speech segregation by examining the across-ear interference effects of three different types of highly intelligible but qualitatively unnatural synthetic speech signals and comparing them to the across-ear interference effects of normal speech. The results are discussed in terms of their implications for human speech segregation.

II. GENERAL METHODOLOGY

All three of the experiments conducted in this study were based on the coordinate response measure (CRM) for multitalker communications research, a call-sign, color, and number-based intelligibility test (Moore, 1981) that is particularly well suited for listening tasks that involve more than one simultaneous speech signal (Moore, 1981; Brungart *et al.*, 2001; Brungart and Simpson, 2002). In a typical trial in the CRM task, a listener is presented with one or more sentences of the form "Ready (call sign) go to (color) (number) now" and asked to identify the color and number combination that was directly addressed to a preassigned "target" call sign (usually "baron"). In this series of experiments, the CRM phrases were drawn from a publicly available corpus (Bolia *et al.*, 2000) that consists of CRM phrases spoken by four male and four female talkers with all possible combinations of eight call signs ("arrow," "baron," "charlie," "eagle," "hopper," "laker," "ringo," "tiger"), four colors ("blue," "green," "red," "white"), and eight numbers (1–8), for a total of 2048 unique sentences.

Two different types of experiments were conducted with each of the three different synthetic speech signals examined in this study. Both involved listeners who were seated at one of three identical Windows-based PC computers located in three different quiet listening rooms. The first type of experiment was a straightforward single-talker listening experiment that examined the overall intelligibility of the different synthetic CRM speech signals. In each trial of these intelligibility experiments, a target phrase was randomly selected from all the available synthetic phrases containing the target call sign "baron," scaled to a comfortable listening level (roughly 70 dB SPL), and presented to the listener over headphones (AKG240) through a 24-bit sound card (Creative Labs Audigy). The listener's task was simply to use the computer mouse to select the color and number combination contained in the stimulus from a grid of colored digits displayed on the CRT of the control computer.

The second type of experiment was a replication of the dichotic CRM listening task first used by Brungart and Simpson (2002). In each trial of this task, the signal presented to the right (target) ear always consisted of a mixture of two simultaneous phrases from the unprocessed natural-speech CRM corpus: a target phrase, which was randomly selected from the phrases containing the call sign "baron," and a masking phrase, which was randomly selected from all the phrases spoken by a different same-sex talker that contained a different call sign, color, and number from the target phrase. The rms level of the target phrase was also scaled relative to the masking phrase to produce one of five different signal-to-noise ratios (−8, −4, 0, 4, or 8 dB).

The signal presented to the left (unattended) ear consisted of (a) silence; (b) a second masking phrase randomly selected from all the phrases in the standard CRM corpus spoken by a different talker of the same sex as the target talker that contained a different call sign, color, and number than either of the two phrases in the target ear; or (c) a synthetic CRM speech signal that was generated according to the procedures outlined in the following sections.

The participants in this dichotic CRM task were instructed to listen in the right ear for the target phrase containing the call sign "baron" and respond by selecting the color and number coordinates contained in that target phrase from the array of colored digits displayed on the screen of the control computer.

The next three sections describe how these experiments were implemented with the three different types of synthetic speech signals that were examined in this investigation of dichotic cocktail-party listening.

III. MODULATED NOISE-BAND SPEECH

One example of a stimulus that is qualitatively much different from speech but still highly intelligible is modulated noise-band (MNB) speech. MNB speech consists of fixed-frequency bands of noise that are independently amplitude modulated to match the envelopes of the corresponding frequency regions in an arbitrary target speech signal (Shannon *et al.*, 1995). When MNB speech is generated from a relatively large number of independently modulated bands of noise, it closely resembles whispered or unvoiced speech. However, as the number of modulated bands is reduced, the spectral detail in the target speech signal is lost and the MNB speech becomes progressively less similar to normal speech. Previous research has shown that MNB speech produces near-perfect vowel intelligibility with eight or more frequency bands, and near-perfect sentence intelligibility with five or more frequency bands (Dorman *et al.*, 1997). As the number of bands is reduced below five, intelligibility systematically decreases until it approaches chance performance in the one-band case where the stimulus is reduced to an amplitude-modulated broadband noise.

As discussed earlier, previous experiments have shown that continuous noise produces little or no across-ear interference in dichotic listening, but that speech does. Because MNB speech systematically changes from a qualitatively noise-like stimulus to a more speech-like stimulus as the number of frequency bands increases, one might also expect the number of frequency bands in MNB speech to influence the amount of across-ear interference it causes in dichotic listening. Experiment 1 was conducted to test this hypothesis. The experiment was divided into two parts. Experiment 1a examined MNB speech intelligibility as a function of the number of independently modulated frequency bands in the stimulus. Experiment 1b examined the contralateral interference effects these MNB stimuli caused in a dichotic cocktail-party listening task.

A. Experiment 1a: Intelligibility

1. Methods

a. Listeners. Nine paid volunteer listeners (four male and five female) participated in the experiment. All had clinically normal hearing (thresholds less than 15 dB HL from 500 Hz to 8 kHz), and their ages ranged from 19–53 years. All of the listeners had participated in previous experiments that utilized the speech materials used in this study.

b. MNB speech materials. For the purposes of this study, only a subset of the standard CRM corpus was processed to generate MNB speech. This subset consisted of all the phrases containing the call signs "tiger," "eagle," and

TABLE I. Cutoff frequencies (in kHz) of the independent frequency bands used to generate the MNB speech in experiment 1.

Number of bands	Start	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.05	4.00														
2	0.05	0.86	4.00													
3	0.05	0.47	1.45	4.00												
5	0.05	0.26	0.61	1.18	2.17	4.00										
10	0.05	0.14	0.26	0.41	0.61	0.86	1.18	1.61	2.17	2.94	4.00					
15	0.05	0.11	0.18	0.26	0.35	0.47	0.61	0.77	0.96	1.18	1.45	1.78	2.17	2.66	3.25	4.00

"baron" spoken by two male talkers (talkers 2 and 3 from the corpus) and two female talkers (talkers 6 and 7 from the corpus), for a total of 384 phrases.

The phrases were converted to MNB stimuli with the PRAAT speech processing software package (Boersma, 1993). The phrases were first downsampled to 20 kHz and low-pass filtered at 4 kHz. They were then converted into the frequency domain with an FFT, divided into the required number of subbands,[1] and converted back in the time domain where the intensity contours of each subband were extracted by squaring the signals and convolving them with a 64-ms Kaiser window. A pink-noise excitation signal was then converted into the frequency domain, divided into the same number of subbands as the speech stimulus, and converted back into the time domain. Each subband of this noise stimulus was amplitude modulated with the intensity contour extracted from the corresponding subband of the speech signal, and the resulting amplitude-modulated noise bands were added together to construct the final MNB speech signal.

Six different MNB stimuli were generated for each phrase in the reduced corpus, each with a different number of independently modulated frequency bands (1, 2, 3, 5, 10, and 15). Thus, a total of 2304 sentences was available for use in the experiment. Note that the frequency bands were equally spaced on an ERB scale in the range from 50 Hz to 4 kHz, as illustrated in Table I.

c. Procedure. The experiment was conducted according to the procedures for CRM intelligibility testing outlined in Sec. II. The data collection was divided into six blocks of 60 trials, with each block containing ten trials for each of the six possible numbers of bands in the MNB corpus (1, 2, 3, 5, 10, or 15). Thus, each listener participated in a total of 60 trials for each number of bands tested in the experiment.

2. Results and discussion

The results of experiment 1a are shown in the left panel of Fig. 1. The intelligibility of the MNB speech increased systematically from around 15% to near 100% as the number of bands increased from one to five. For comparison, we have also replotted the results for the two speech corpora (out of a total of five tested) that produced the best and worst overall performance in Dorman *et al.*'s (1997) evaluation of the intelligibility of MNB speech: the Iowa Consonant Test of 16 consonants in an /aCa/ format spoken by a single male talker (Tyler *et al.*, 1986) [which was also the speech corpus used in the earlier study by Shannon (1995)]; and a multitalker vowel intelligibility test comprised of the 11 vowels in the words "hawed, heed, hid, hayed, head, had, hod, hood, hoed, who'd, and heard" spoken by three men, three women, and three girls (Hillenbrand *et al.*, 1995). These results show that the intelligibility levels obtained with the CRM corpus used in this experiment were roughly comparable to those reported for the relatively easy Iowa Consonant Test used in earlier MNB experiments by Shannon (1995) and Dorman *et al.* (1997).

B. Experiment 1b: Across-ear interference

1. Methods

a. Listeners. The same nine listeners who participated in experiment 1a also participated in experiment 1b.

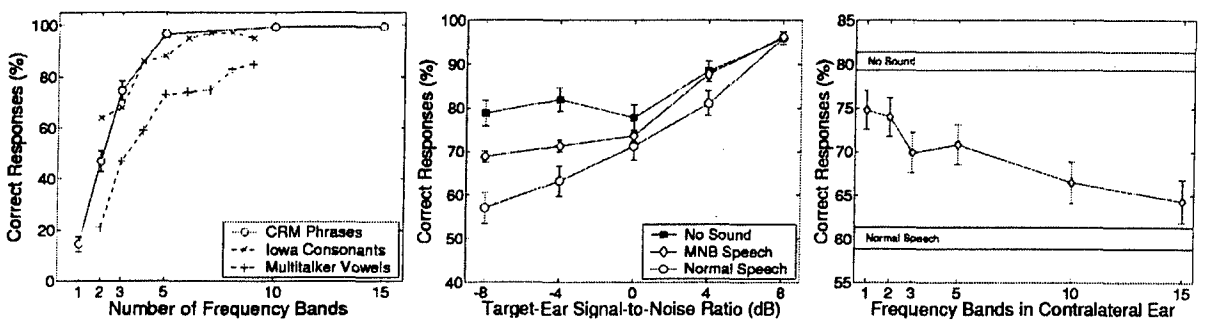


FIG. 1. The open circles in the left panel show the percentage of trials in which the listeners correctly identified both the color and number coordinates in experiment 1a, which measured speech intelligibility as a function of the number of frequency bands in the MNB speech stimuli. For comparision, the results obtained by Dorman *et al.* (1997) for similarly processed Iowa consonants and multitalker vowels have also been replotted in this panel. The center panel shows the percentage of correct color and number identifications in experiment 1b as a function of target-ear SNR. The black squares and open circles show performance in the control conditions where there was no contralateral masker or a normal-speech masker. The shaded diamonds show performance averaged across all the conditions with a contralateral MNB speech masker. The right panel shows the percentage of correct color and number identifications in the negative target-ear SNR conditions of experiment 1b. The shaded bars in that panel show mean performance ±1 standard error in the no-sound and normal-speech control conditions. The error bars represent the 95% confidence intervals for each data point.

b. Procedure. The experiment was conducted according to the procedures for the dichotic CRM task outlined in Sec. II. In the conditions where the masking phrase presented in the left ear consisted of synthetic speech, that masking phrase was randomly selected from the MNB-processed CRM phrases that contained a different call sign, color, and number than either of the two phrases in the target ear.[2] When the normal speech phrase was used in the unattended ear, it was low-pass filtered to 4 kHz to match the bandwidth of the MNB-processed speech stimuli and then scaled to match the rms level of the masking talker in the target ear. When the MNB speech was used in the unattended ear, it was also scaled to match the overall rms level of the masking talker in the target ear.

The data collection was divided into 40 blocks of 80 trials, with two repetitions of each of the eight possible contralateral masking conditions (silence, normal speech, or 1-, 2-, 3-, 5-, 10-, or 15-band MNB speech) at each of the five target-ear SNR values in each block. Thus, each of the nine listeners participated in a total of 80 trials for each combination of contralateral masker and target-ear SNR tested in the experiment.

2. Results and discussion

The results of the experiment are shown in the middle and right-hand panels of Fig. 1. The middle panel shows performance as a function of the SNR in the target ear for the conditions with no sound, MNB speech, or normal speech in the contralateral ear. For simplicity, all of the different MNB conditions have been averaged together to create the middle curve in the panel. In the no-sound and normal-speech control conditions, the results were similar to those in an earlier experiment that used the same CRM stimuli and the same dichotic listening task used in this experiment (Brungart and Simpson, 2002). In the condition with no contralateral masker (black squares), performance increased as the SNR increased above 0 dB, but plateaued at approximately 80% correct responses for SNR values at or below 0 dB. In the condition with a normal speech contralateral masker (open circles), performance was similar to the no-sound condition when the SNR was +8 dB, but it decreased much more rapidly with decreasing SNR. As a consequence, performance at −8-dB SNR was roughly 20 percentage points worse with a contralateral speech masker than it was with no contralateral masking signal. The gray diamonds show performance averaged across the six MNB speech conditions of the experiment. As we hypothesized, the results for the MNB speech consistently fell between those for the no-sound and normal-speech contralateral masking conditions. This suggests that MNB speech causes more contralateral interference than no masker, but less interference than a normal speech masker.

The middle panel of Fig. 1 also indicates that the contralateral maskers had the greatest impact on performance when the target-ear SNR was less than 0 dB. Consequently, the right panel of Fig. 1 focuses on the differences between the MNB-speech conditions in trials where the target-ear SNR was negative. For comparison, shaded regions of the figure show mean performance ±1 standard error in the no-sound and normal-speech control conditions of the experiment. These results show that there was indeed a systematic decrease in performance as the number of frequency bands in the MNB speech increased. A one-factor within-subjects ANOVA on the arcsine-transformed results of the individual subjects for each of the eight contralateral masking conditions (no sound, normal speech, and 1-, 2-, 3-, 5-, 10-, or 15-band MNB speech) indicated that this effect was statistically significant ($F_{(7,56)}=15.58$, $p<0.0001$), and a subsequent *post hoc* test (Fisher LSD, $p<0.05$) indicated the following significant results:

(1) All the MNB speech conditions were significantly worse than the no-sound control condition.
(2) All the MNB speech conditions except the 15-band condition were significantly better than the normal-speech control condition.

Thus, it seems that even the single-band MNB speech distractor, which scored only slightly better than chance in the intelligibility test in experiment 1a, produced a significant amount of across-ear interference in the dichotic listening task of experiment 1b. As the number of frequency bands increased, so did the across-ear interference caused by the MNB speech. However, the amount of interference did not plateau at the 5-band level where intelligibility reached near 100% performance in experiment 1a. Rather, it continued to increase until the 15-band point, where the MNB speech was producing nearly as much contralateral interference as normal speech.

IV. MODULATED SINE-BAND SPEECH

Modulated noise-band speech is qualitatively much different from normal voiced speech, but when it consists of a large number of frequency channels it can sound similar to whispered or unvoiced speech. Thus, it is conceivable that the increase in across-ear masking that occurred in the 15-band condition of experiment 1 could be directly related to the similarity of the speech in that condition to natural whispered speech. It is possible, however, to generate a stimulus that contains the spectral information similar to MNB speech but sounds unnatural even when it contains a large number of frequency channels. This speech is generated by replacing the amplitude-modulated noise bands in MNB speech with amplitude-modulated sine waves fixed at the center frequencies of those bands. Previous experiments that have compared this type of modulated sine-band (MSB) speech to MNB speech have found very little difference in intelligibility between the two types of simulated speech (Dorman *et al.*, 1997), despite the large qualitative difference between the two types of speech signals. Experiment 2 was conducted to evaluate the amount of across-ear interference generated by MSB speech in a dichotic cocktail-party listening task.

A. Experiment 2a: Intelligibility

1. Methods

a. Listeners. Eight paid volunteer listeners with clinically normal hearing (five male and three female) participated in the experiment. Six of the listeners were also participants in experiments 1a and 1b.

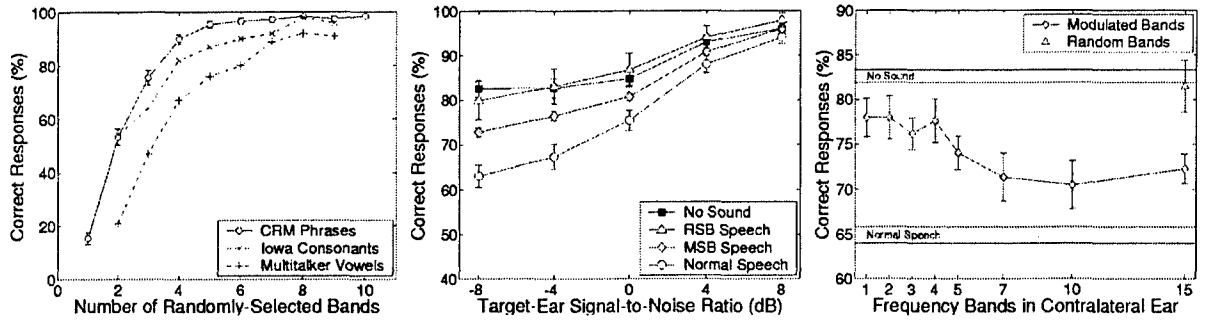


FIG. 2. The left panel shows the percentage of correct color and number identifications in experiment 2a, which measured speech intelligibility as a function of the number of frequency bands in the MSB speech stimuli. As in Fig. 1, the intelligibility results obtained by Dorman *et al.* (1997) for MSB-processed Iowa consonants and multitalker vowels have been replotted in this panel for comparison. The center panel shows the percentage of correct color and number identifications in experiment 2b as a function of target-ear SNR. The black squares and open circles show performance in the control conditions where there was no signal in the contralateral ear (squares) or a normal-speech signal in the contralateral ear (circles). The shaded diamonds show performance averaged across all the conditions with a contralateral MNB speech masker. The right panel shows the percentage of correct color and number identifications in the negative target-ear SNR conditions of experiment 1b. The shaded bars in that panel show mean performance ±1 standard error in the no-sound and normal-talker control conditions. The error bars represent the 95% confidence intervals for each data point.

b. Speech materials. The MSB speech stimuli were derived from the male-talker sentences from the same CRM speech corpus used in experiment 1.[3] These stimuli were processed with a technique that Arbogast *et al.* (2002) adapted from cochlear implant simulation software originally developed by the House Ear Institute. The sentences in the CRM corpus were first downsampled from 40 to 20 kHz. Then, they were high-pass filtered at 1200 Hz with a first-order Butterworth filter and processed with a bank of 15 fourth-order 1/3rd-octave Butterworth filters with logarithmically spaced center frequencies ranging from 215 to 4891 Hz with a ratio of successive center frequencies of 1.25. The envelopes of each of these channels were extracted by half-wave rectifying the bandpass-filtered signals and low-pass filtering them at 50 Hz. Then, these envelopes were used to modulate pure tones with zero starting phases and center frequencies at the midpoints of each filter band. Individual sound files were created for each of these 15 bands for the 256 CRM phrases spoken by each of the male talkers in the CRM corpus, and the stimuli used in the experiment were generated by randomly selecting 1–10 of these individual bands from the same original CRM phrase and summing them together electronically.[4]

c. Procedure. Other than the method used to generate the speech stimuli, the experimental procedure was essentially identical to the one used in experiment 1a. Each block of trials in the experiment consisted of 12 repetitions of each of the 10 MSB speech conditions of the experiment (i.e., 1–10 individual randomly selected bands). Each listener participated in 10 blocks of trials, so a total of 960 trials was collected in each of the 10 conditions of the experiment (8 listeners×10 blocks×12 repetitions).

2. Results and discussion

The left panel of Fig. 2 shows the intelligibility results from experiment 2a. Intelligibility was poor (<20%) in the one-band condition, but it increased systematically with the number of bands, plateauing at near 100% performance when five independent frequency bands were present in the stimulus. Overall, this performance function is very similar to the one obtained with the MNB-processed CRM stimuli in experiment 1a (plotted in the left panel of Fig. 1). The intelligibility scores were, however, slightly higher than those reported for the MSB-processed Iowa consonants in the earlier experiment by Dorman *et al.* (1997), which have been replotted in the figure for comparison. Comparing Figs. 1 and 2, it is apparent that the CRM stimuli used in this experiment produced intelligibility levels that were very similar to those obtained for the Iowa consonants in the MNB processing condition, but somewhat better than those obtained for the Iowa consonants in the MSB condition. This difference may, in part, be due to the fact that Dorman and his colleagues generated their MSB stimuli with modulated sinewave bands that were always evenly distributed across the speech spectrum, while the stimuli in this experiment were generated with modulated sinewave bands that were randomly selected from the 1/3rd-octave bands that were available in the 15-band MSB processed speech. The difference might also simply be due to the semantic differences between the two speech corpora. In either case, the results shown in Fig. 2 indicate that the random-frequency MSB speech used in experiment 2a produced intelligibility in the CRM task that was comparable to that obtained for MNB speech generated with the same number of frequency bands in experiment 1a.

B. Experiment 2b: Across-ear interference

1. Methods

a. Listeners. Seven of the eight listeners who participated in experiment 2a also participated in experiment 2b.

b. Speech materials. The MSB conditions of experiment 2b used the same stimulus processing as described in experiment 2a. In addition to these MSB speech conditions, a 15-band random sine-band (RSB) speech control condition was also tested. The RSB speech was produced by randomizing the phase component of a standard 15-band MSB speech signal. This was accomplished by multiplying the long-term complex spectrum (FFT) of a randomly selected 15-band MSB speech signal by the long-term complex spectrum of a broadband Gaussian noise and taking the inverse FFT of this

multiplied frequency-domain signal (Arbogast *et al.*, 2002). This processing resulted in an unintelligible waveform that was spectrally identical to the MSB speech but contained no phonetic information about the original utterance.

c. Procedure. Experiment 2b used the same dichotic CRM task used in experiment 1b, with the exception that only two of the talkers were used as target talkers (the male talker 1 and the female talker 6) with the same target talker used in every stimulus presentation within the same block of trials. In the conditions where the masking phrase presented in the left ear consisted of synthetic speech, that masking phrase consisted of MSB speech with 1, 2, 3, 5, 7, 10, or 15 bands or RSB speech with 15 bands. In all cases, the masking speech signal was selected to have a different color and number than either of the two phrases in the target ear.

The data collection was divided into blocks of approximately 70 trials with each subject participating in roughly 100 blocks, for a total of 6864 trials per subject or 48 048 trials in the experiment. All subjects participated in all conditions, and the total number of trials per condition ranged from 1698 trials for the 15-band RSB speech condition to 7305 trials for the 15-band MSB speech condition.

2. Results and discussion

The middle and right-hand panels of Fig. 2 show the overall results of experiment 2b. The middle panel shows performance as a function of the SNR in the target ear for the conditions with no sound, RSB speech, MSB speech, or normal speech in the contralateral ear. Again, the different MSB conditions have been averaged together to simplify the visual presentation of the data in this panel. The results show that the no-sound (black squares) and normal-speech (open circles) control conditions were essentially identical to the corresponding conditions of experiment 1b (shown in Fig. 1). Also, as with the MNB speech in experiment 1b, the results with the MSB speech in experiment 2b consistently fell between these two control conditions. In contrast, performance with the 15-band RSB speech (open triangles) was essentially identical to the no-sound control condition.

The right panel of Fig. 2 shows performance in the different MSB-speech conditions averaged across trials where the target-ear SNR was less than 0 dB. Again, the shaded regions of the figure show mean performance ± 1 standard error in the no-sound and normal-speech control conditions of the experiment. Performance in the 15-band RSB condition is also shown by the white triangle. The arcsine-transformed data from the individual subjects in each of the 11 contralateral masking conditions (no sound, normal speech, 1-, 2-, 3-, 4-, 5-, 7-, 10-, or 15-band MSB speech or RSB speech) were also subjected to a one-factor within-subjects ANOVA, which indicated that the main effect of the contralateral masking condition was statistically significant ($F_{(10,60)}=9.52$, $p<0.0001$). A subsequent *post hoc* test (Fisher LSD, $p<0.05$) revealed the following significant effects:

(1) All the MSB speech conditions except the 1-band condition were significantly worse than the no-sound control condition.[5]
(2) All the MSB speech conditions were significantly better than the normal-speech control condition.
(3) The 1-, 2-, and 4-band conditions were significantly better than the 7-, 10-, and 15-band conditions.
(4) There was no significant difference between the 15-band RSB condition and the no-audio control condition.

As in the MNB condition, the results show a general trend of increasing across-ear interference with an increasing number of frequency bands. However, in the limiting 15-band case, performance appeared to be slightly better relative to the normal-speech control condition with MSB speech. This may reflect the fact that 15-band MNB speech sounds similar to natural whispered speech, while MSB speech sounds decidedly unnatural even with 15 frequency bands.

It is also interesting to note that the RSB speech failed to produce any measurable across-ear interference even though it contained all 15 possible frequency bands. The long-term magnitude spectrum of this RSB speech signal was identical to that of the 15-band MSB speech, so it seems that the across-ear interference caused by the MSB speech cannot be explained by spectral content alone. Rather, it seems that the speech-like temporal modulations in the individual bands of the MSB speech were critical to the across-ear interference effects that occurred with those stimuli. This seems to be consistent with our earlier finding that the contralateral noise that was shaped to match the long-term rms spectrum of CRM speech produced little or no across-ear interference in the dichotic CRM task (Brungart and Simpson, 2002). It is also consistent with the results of Arbogast *et al.* (2002), who also found a substantial difference between the masking properties of MSB and RSB speech in normal binaural listening environments. In their experiment, they randomly selected 8 of the 15 bands for use in the target speech signal, and allocated 6 of the remaining bands either to an MSB speech masker or an RSB speech masker. Their results showed that the speech reception threshold (SRT) was 22 dB lower with RSB masking speech than it was with MSB masking speech, presumably because the speech-like MSB masker was more easily confused with the target speech signal. Our results show that this masking difference between MSB and RSB speech extends to the case where the target and masking speech signals are presented to different ears.

V. SINEWAVE SPEECH

An additional type of "speech-like" stimulus that is qualitatively different from speech but still highly intelligible is so-called "sinewave speech," which consists of a small number of time-varying amplitude-modulated sine waves that track the formant frequencies of a speech signal (Remez *et al.*, 1981). Experiment 3 was conducted to determine whether this kind of stimulus also produces across-ear interference in a dichotic cocktail-party listening task.

A. Experiment 3a: Intelligibility

1. Methods

a. Listeners. Nine paid volunteer listeners with clinically normal hearing (four male and five female) participated

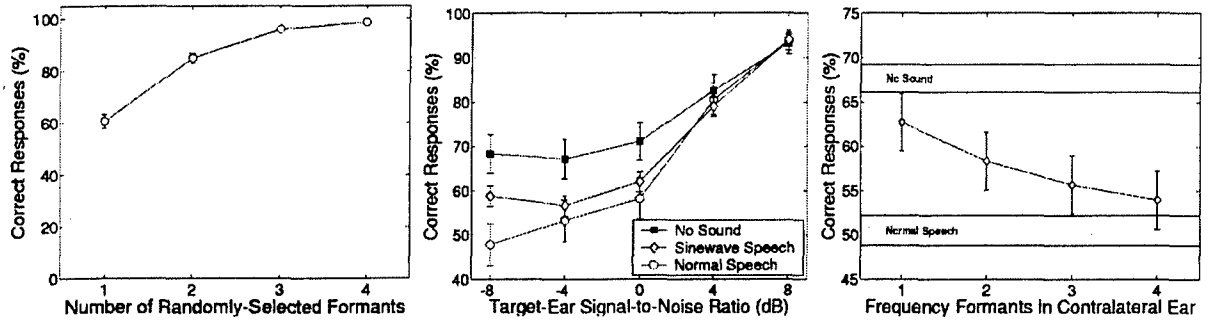


FIG. 3. The left panel shows the percentage of correct color and number identifications in experiment 3a, which measured speech intelligibility as a function of the number of formants in the sinewave speech stimuli. The center panel shows the percentage of correct color and number identifications in experiment 3b as a function of target-ear SNR. The right panel shows the percentage of correct color and number identifications in the negative target-ear SNR conditions of experiment 3b, which measured the effects of a contralateral sinewave speech interferer on two-talker segregation performance in the listener's right ear. The error bars represent the 95% confidence intervals for each data point.

in the experiment. Six of the nine listeners had previously participated in experiment 1, and four had previously participated in experiment 2.

b. Speech Materials. The sinewave speech stimuli were processed directly from the CRM speech corpus using LPC-based MATLAB script files that have been made publicly available on the Internet by Ellis (2003). These scripts estimate the magnitudes and frequencies of the first four formants in each 2.6-ms frame from the filter pole positions derived from an LPC analysis. The CRM sentences were resampled to an 8-kHz rate prior to performing this LPC analysis, resynthesized into sinewave speech, and then resampled to a 50-kHz rate prior to presentation to the listeners. This processing was done in real time within each trial of the experiment.

c. Procedure. Again, the procedure used in experiment 3a was essentially identical to the one used in experiments 1a and 2a. In each trial of the experiment, a target phrase was randomly selected from all the phrases containing the target call sign "baron" in the CRM corpus. This target phrase was processed into sinewave speech, and then one, two, three, or four of the first four formants were randomly selected for inclusion in the stimulus. The data collection was divided into 10 blocks of 60 trials, with each block containing 15 trials for each of the four possible numbers of formants (1, 2, 3, or 4). Thus, each listener participated in a total of 600 trials in the experiment.

2. Results and discussion

The left panel of Fig. 3 shows the intelligibility results from experiment 3a. The major difference between these results and the earlier results with the MNB and MSB speech signals in experiments 1a and 2a is the much higher intelligibility score that was achieved with just a single randomly selected formant (near 60%, versus less than 20% for the other two stimulus types). This reflects the fact that the sinewave speech adapts itself to track variations in the frequencies of the formants, while the MSB and MNB stimuli provide spectral information only in fixed frequency regions. Note that intelligibility approaches 100% for sinewave speech stimuli comprised of three or more randomly selected formants.

B. Experiment 3b: Across-ear interference

1. Methods

a. Listeners. The same nine listeners who participated in experiment 3a also participated in experiment 3b.

b. Procedure. The procedure used in experiment 3b was essentially identical to the one used in experiment 1b. When a synthetic speech signal was presented in the left ear, it consisted of sinewave speech that was generated with 1, 2, 3, or 4 randomly selected formants using the procedure outlined in the previous section. When a natural speech phrase was presented in the unattended ear, it was low-pass filtered to 4 kHz to match the maximum bandwidth of the sinewave speech stimuli. In all cases, the interfering speech signal in the contralateral ear was scaled to match the rms level of the masking talker in the target ear.

The data collection was divided into 24 blocks of 60 trials, with two repetitions of each of the six possible contralateral masking conditions (silence, normal speech, or 1-, 2-, 3-, or 4-band sinewave speech) at each of the five target-ear SNR values in each block. Thus, each of the nine listeners participated in 1440 trials in the experiment, for a total of 432 trials for each combination of target-ear SNR and contralateral-ear masker tested in the experiment.

2. Results and discussion

The results of experiment 3b are shown in the right two panels of Fig. 3. The middle panel of the figure shows performance as a function of the target-ear SNR. Again, the four sinewave-speech conditions have been averaged together into a single curve (shaded diamonds) to allow an easy comparison to the no-sound (black circles) and normal-speech (open circle) control conditions. Although performance in these control conditions was markedly lower than in experiments 1b and 2b (presumably because of the different mix of subjects), the overall pattern of performance was the same: a plateauing in performance at negative SNR values in the no-sound condition, and a roughly 20-percentage point decrease in performance in the normal-speech condition at an SNR of −8 dB.

Performance with the sinewave-speech contralateral maskers (gray diamonds) again fell between these two control conditions, with the largest decrease relative to the no-

sound condition occurring at negative target-ear SNR values. The right panel of Fig. 3 shows performance as a function of the number of formant frequencies in the contralateral sinewave speech masker averaged across trials where the target-ear SNR was less than 0 dB. As before, the arcsine-transformed data from the individual subjects in each of the six contralateral masking conditions (no sound, normal speech, and 1-, 2-, 3-, or 4-formant sinewave speech) were analyzed by a within-subjects ANOVA, which indicated that the main effect of the contralateral masking condition was statistically significant ($F_{(5,40)}=11.13$, $p<0.0001$). A subsequent *post hoc* test (Fisher LSD, $p<0.05$) found the following significant differences:

(1) All the sinewave speech conditions except the 1-formant condition were significantly worse than the no-sound control condition.
(2) All the sinewave speech conditions except the 4-formant condition were significantly better than the normal-speech control condition.
(3) Performance in the 1-formant condition was significantly better than the 3- and 4-formant conditions.

Thus, we see that, as with the other types of simulated speech signals tested in these experiments, sinewave speech tends to produce more across-ear interference than noise in dichotic listening, but less interference than normal speech. Also, the data suggest that sinewave speech may be somewhat more efficient at generating across-ear interference than MSB or MNB speech. Sinewave speech produced almost as much interference as normal speech with just 4 formant frequency bands, a level of interference that required 15 bands for the MNB speech and never occurred with the MSB speech. However, it should be noted that, like MSB speech, the sinewave speech stimuli never sounded remotely similar to any type of natural speech even with the largest number of frequency bands tested. Thus, it seems that the difference in across-ear interference that occurred between experiment 1b and experiment 2b cannot be accounted for solely by the whisper-like characteristics of MNB speech when it contains a large number of frequency bands.

VI. GENERAL DISCUSSION

This paper has presented the results of three experiments comparing the across-ear interference generated by three distinct types of simulated speech to the amount of across-ear interference that occurs with a normal speech signal. Although the three types of simulated speech were qualitatively much different, their contralateral masking characteristics were similar: (1) all produced some amount of contralateral interference when they contained only one or two frequency bands; (2) the amount of contralateral interference increased systematically with the number of frequency bands; and (3) performance for the maximum number of frequency bands tested approached the normal-speech control condition.

The results of the experiments described in this paper, along with those of our earlier study examining the effects of a contralateral masker on dichotic speech perception (Brungart and Simpson, 2002), allow us to answer a number of important questions regarding the across-ear interference that occurs in dichotic cocktail-party listening.

(1) *Is there a threshold level of similarity to speech that must be reached in order for a speech-like signal to generate across-ear interference in a dichotic cocktail-party listening task*? No. With all three of the synthetic speech stimuli we tested, the amount of across-ear interference increased gradually as the number of bands increased. Similarly, in our earlier experiment, there was a gradual decrease in across-ear interference as the speech signal in the contralateral ear was masked with noise (Brungart and Simpson, 2002). This argues against the existence of a "threshold" level of speech-like attributes that must be reached in order for a contralaterally presented speech signal to interfere with speech perception in the opposite ear.
(2) *Is long-term spectral similarity to speech necessary or sufficient for a signal to generate across-ear interference in a dichotic cocktail-party listening task*? No. In our earlier experiment, we showed that Gaussian noise that was spectrally shaped to match the long-term spectrum of speech caused little or no across-ear interference in dichotic listening. In this series of experiments, we demonstrated that at least two types of signals with long-term spectra that differed dramatically from normal speech (MSB speech and sinewave speech) generated substantial amounts of across-ear interference. From these two results, we can conclude that spectral similarity to speech is neither necessary nor sufficient for a sound to produce across-ear interference in dichotic listening. Further evidence for the relatively minor role that long-term spectrum plays in contralateral masking was provided by the results of experiment 2b: the long-term spectrum of the 15-band RSB speech contralateral masker used in that experiment was identical to the spectrum of the 15-band MSB speech, but the RSB speech produced far less contralateral interference than the MSB speech masker. Again, this suggests that overall spectrum is a relatively unimportant parameter in determining the amount of across-ear interference a contralateral masking signal will generate.
(3) *Is intelligibility necessary for a signal to generate across-ear interference in a dichotic cocktail-party listening task*? No. In our earlier experiment, we demonstrated that time-reversed speech produced just as much across-ear interference as normal speech when the target-ear signal-to-noise ratio was less than 0 dB. Thus, it appears that unintelligible signals can produce just as much contralateral interference as intelligible signals in dichotic listening.
(4) *Is intelligibility sufficient for a signal to generate across-ear interference in a dichotic cocktail-party listening task*? Probably. We have not tested all of the synthetic signals that could conceivably be used to generate intelligible speech, but we have examined three of the least speech-like signals that have been demonstrated to contain usable verbal information, and we have shown that

all three produce significant amounts of across-ear interference in dichotic listening. This leads us to suspect that any signal capable of conveying the useful phonetic information contained in normal speech will produce some across-ear interference in a dichotic cocktail-party task. However, we should point out that, to this point, we have only tested signals that have been gated on and off at approximately the same time as the target speech. It is possible that adaptation might allow listeners to perform better in the dichotic listening task if the contralateral masker were a continuous speech signal that was turned on some time before the onset of the target speech.

(5) *Are speech-like temporal modulations in the spectral envelope of a signal sufficient to generate across-ear interference in a dichotic cocktail-party listening task*? Yes. The MNB speech differed from broadband speech-shaped noise only in terms of the introduction of speech-like modulations in the spectral envelope, and these modulations were sufficient to generate a substantial amount of contralateral interference in the dichotic listening task. Similarly, the MSB speech differed from the RSB speech only in terms of its envelope modulations, and these modulations were sufficient to generate a substantial amount of across-ear interference. However, it is important to note that the modulations that appear to be most critical to the across-ear interference effects demonstrated in these experiments are the varying narrowband temporal modulations that occur in speech, and that the contralateral masking effects of these modulations are probably limited to listening tasks where the target signal is also speech-like. Listening tasks involving nonspeech target signals may be more sensitive to contralateral interference from signals with different qualitative characteristics and different modulation patterns. Kidd *et al.* (2003), for example, examined performance in a nonspeech dichotic listening task that required listeners to detect fixed-frequency pulsed tone targets in the presence of tone or noise maskers and found that contralaterally presented fixed-frequency tone complexes that were coherently gated with the target produced significant amounts of across-ear interference, but that contralaterally presented notch-filtered noise that was coherently gated with the target did not. Thus, in that case, significant across-ear interference only occurred when the contralateral masking signal was synchronously gated with *and* qualitatively similar to the target signal. Consequently, it is likely that the contralaterally presented synthetic speech signals that caused significant across-ear interference in this experiment would have little or no effect on performance in the dichotic tone-detection task examined by Kidd and his colleagues. Thus, while speech-like modulations appear to be sufficient to produce across-ear interference in dichotic speech perception tasks, other factors—such as qualitative target–masker similarity—can strongly influence the across-ear interference effects that occur in other kinds of listening tasks.

(6) *Are speech-like temporal modulations in the spectral envelope of a signal necessary to generate across-ear interference in a dichotic cocktail-party listening task*? Possibly. We have not yet tested any signals that generate a substantial amount of across-ear interference and do not have speech-like temporal envelope fluctuations. Thus, while we cannot rule out the possibility that such signals exist, we do not yet have any evidence to demonstrate that signals without speech-like envelope fluctuations can cause across-ear interference in dichotic cocktail-party listening.

(7) *What are the requirements for modulations in the spectral envelope of a signal to be "speech-like" in the sense that they will produce significant amounts of across-ear interference in a dichotic cocktail-party listening task*? This is perhaps the most interesting remaining research question related to the contralateral interference effects we have demonstrated in our dichotic listening experiments. All of our experiments to this point suggest that certain types of contralaterally presented audio signals are identified as "speech-like" by some preattentive central auditory processing mechanism, and that signals that fall into this category interfere with a listener's ability to segregate speech signals presented in the opposite ear. The results of this experiment strongly suggest that speech-like modulations in the spectral envelope play an important role in determining what kinds of signals are identified as speech-like by this central processing. Furthermore, our earlier results have shown that these speech-like fluctuations do not necessarily have to be intelligible to cause interference: time-reversed speech, which is unintelligible but has envelope fluctuations similar to those in normal speech, produces nearly as much across-ear interference as normal speech. At this point, however, it is not clear what the parameters are that determine whether or not these envelope fluctuations are speech-like. What range of modulation frequencies will generate this type of interference? Do the modulation frequencies have to vary over time like they do in natural speech, or will constant envelope modulations cause the same amount of contralateral interference? Do the modulations have to be correlated across frequency as they are in natural speech, or do independent speech-like envelope modulations (such as those that would occur with a stimulus matching the envelopes of different utterances at different frequency regions) also interfere? The answers to these questions are important, because they have the potential to provide valuable insights into the processing methods that listeners unconsciously use to segregate complicated auditory scenes containing more than one simultaneous speech signal. This information might also provide some new ideas about how to produce machine listening devices capable of segregating multiple-talker listening environments using the same strategies that human listeners use for these segregation tasks. At this point, however, only further research can provide the answers to these important questions about dichotic speech perception.

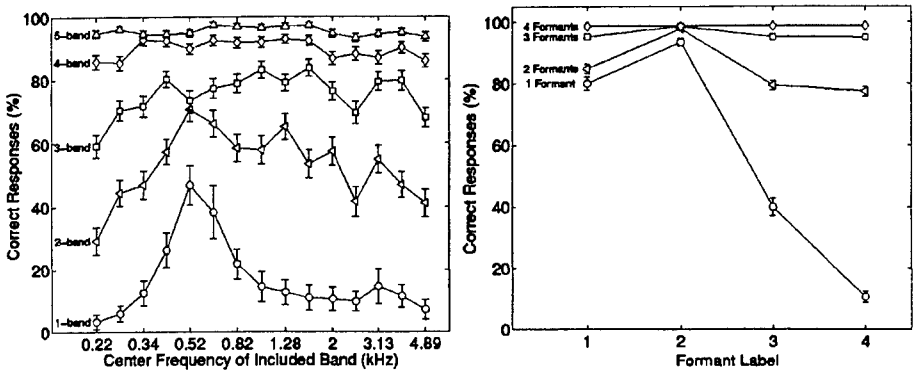


FIG. 4. Each curve in the left panel shows the percentage of correct responses in experiment 2a averaged across all the stimuli with the same number of frequency components that contained the indicated frequency band. Similarly, the curves in the right panel show the percentage of correct responses in experiment 3a averaged across all the stimuli with the same number of frequency components that contained the indicated (labeled) formant. The error bars represent ±1 standard error around each data point.

VII. SUMMARY AND CONCLUSIONS

In this series of three experiments, we have demonstrated that three subjectively very different types of synthetic speech signals (MNB speech, MSB speech, and sinewave speech) have similar effects on speech intelligibility when they are presented to the unattended ear in a dichotic cocktail-party listening task. In all three cases, there was a systematic decrease in performance in the two-talker target-ear listening task when the number of frequency bands in the contralateral speech-like masker increased. These results suggest that speech-like fluctuations in the spectral envelope of a signal play an important role in determining the amount of across-ear interference that signal will produce in a dichotic cocktail-party listening task.

In closing, it is perhaps useful to take a step back and consider how this finding relates to our more general understanding of how listeners process multiple simultaneous speech signals in real-world cocktail party listening environments. Clearly, the stimuli examined in this experiment are artificial in the sense that they would never occur in real-world listening. Indeed, even the more general realm of dichotic listening is somewhat unrealistic, because real-world speech signals are almost always perceived binaurally rather than monaurally. However, what these results do allow us to do is begin to gain some insights into the point at which the auditory system starts to make a distinction between signals that are speech-like and should be processed when the listener is performing a speech perception task and those that are "noise-like" and should be discarded. In the long term, these insights might also help us understand the acoustic features that make it difficult for listeners to segregate simultaneously presented speech signals that, from a purely acoustic standpoint, should individually be clearly audible [a concept sometimes referred to as informational masking (Kidd *et al.*, 1998; Freyman *et al.*, 2001, 1999; Brungart, 2001b)]. Further research is now needed to fully examine the relationship between the temporal fluctuations that occur in the envelopes of a speech-like masking signal and the amount of masking such a signal will produce when it is presented in the unattended ear in a dichotic cocktail-party listening task, and to determine the extent to which a similar kind of interference might occur in more realistic binaural cocktail-party listening tasks that more accurately represent the difficulties listeners encounter in real-world verbal communication.

ACKNOWLEDGMENTS

Portions of this research were supported by AFOSR Grant 01-HE-01-COR and NIH/NICDC Grants DC00100, DC045045, and DC04663.

APPENDIX: FREQUENCY WEIGHTING WITH MSB AND SINEWAVE SPEECH

In speech perception, different frequency regions vary in terms of their relative contributions to overall intelligibility. This attribute of speech perception is one of the foundations of the Articulation Index (AI), which assigns different weights to each 1/3rd-octave band to account for differences in the relative importance of each band in the perception of phonetically balanced speech (French and Steinberg, 1947). However, the coordinate response measure speech materials used in these experiments are not phonetically balanced, so their frequency-dependent intelligibility characteristics may differ from those that would ordinarily occur with traditional speech perception tasks (Brungart, 2001a). Thus, it may be useful to analyze the results of experiments 2a and 3a to examine the contributions that different frequency regions made to the overall perception of the CRM stimuli.

Figure 4 shows how performance varied across the possible frequency component combinations that could occur with the MSB stimuli in experiment 2a and with the sinewave speech stimuli in experiment 3a. In the left panel, each curve represents mean performance across all the MSB speech trials in experiment 2a that contained the indicated number of frequency bands. Within each curve, the data points represent mean performance across all the stimuli with that particular number of bands that contained the frequency component indicated by the abscissa. Thus, in the one-band curve (circles), each data point represents performance in stimulus presentations that contained only the designated frequency component. In the two-band curve (left-pointing triangles), each data point represents mean performance across all the trials that contained the designated band plus one other randomly selected band. And, in the five-band curve, each data point represents mean performance across all the trials that contained the designated band plus four other randomly selected bands.

From the one-band curve, it is immediately apparent that the most important frequency component for overall intelligibility in the CRM task was the modulated sinewave at 520

Hz. In fact, the listeners were able to correctly identify both the color and the number in the stimulus almost half the time when the 520-Hz component was the only frequency component present in the stimulus. In comparison, most of the other frequency bands generated only about 10% correct performance when they were presented in isolation. Interestingly, the 520-Hz band is much lower in frequency than the most highly weighted 1/3rd-octave band in the calculation of the AI, which is centered at 2 kHz (French and Steinberg, 1947). This suggests that the phonetic information in the CRM speech corpus is concentrated in a lower frequency range than the phonetic information in long-term running speech.

As additional frequency bands were added to the MSB CRM stimuli, the specific bands contained in the stimuli became increasingly less important until, in the four-band case, the presence of any particular band no longer had any meaningful impact on the overall intelligibility of the stimuli. This suggests that highly intelligible MSB speech can be generated from any five randomly selected bands out of the 15 frequency bands tested in experiment 2.

The right panel of Fig. 4 shows the relative intelligibility contributions of each of the four formant frequencies tested in the sinewave speech stimuli of experiment 3a. As before, each curve represents a different number of formants, and each data point represents average performance across all the trials with the designated number of sinewave components that happened to contain the formant indicated by the abscissa. Again, these results show that there were large differences across the different formant combinations when the number of frequency components was low. Indeed, the data show that intelligibility was close to 100% when just the second formant was present in the stimulus, but was less than 10% when only the fourth formant was present in the stimulus. The performance variations across the different possible combinations largely disappeared when a second formant was added to the stimulus.

The dramatic variations in performance across the different formant combinations in experiment 3a suggest that perhaps there might also have been a significant variation in the amount of contralateral interference caused by these different sinewave speech stimuli. In order to test this hypothesis, the overall performance level was calculated for each of the 15 possible combinations of 1, 2, 3, or 4 formants in the contralateral masking conditions of experiment 3b, and the correlation coefficient was calculated between the scores in these 15 conditions and the intelligibility scores in the 15 corresponding conditions of experiment 3a. The resulting r value was -0.009, suggesting that intelligibility may be a relatively poor predictor of the across-ear interference that will occur when a speech-like masking signal is presented in the unattended ear in a dichotic cocktail party listening task.

[1]This division was accomplished with the Filter (Pass Hann) command in PRAAT. See the PRAAT documentation (Boersma, 1993) for more details.

[2]The MNB speech could be male or female independent of the sex of the target talker. However, because all voicing information is removed, there is little or no apparent difference between male and female MNB speech.

[3]Again, because all of the voicing information was removed, there was little or no discernible difference between male and female talkers in the MSB speech.

[4]In comparing this technique to the one used to produce the stimuli in the earlier study by Dorman *et al.* (1997), it is important to note that this processing technique involves the exclusion of some speech envelope information when the number of bands is reduced, while with Dorman's technique the speech envelope information is not excluded but rather averaged over a larger bandwidth when the number of bands decreases.

[5]Note that on the surface this seems to contrast with the mean results shown in Fig. 2, which show a mean for the 1-band condition that is comparable to the 2-band condition with a slightly smaller error bar. However, a *post hoc* LSD test on the arcsine-transformed data indicates that the 1-band MSB speech signal is not significantly different from the no-sound control condition ($p=0.0615$). The discrepancy reflects the fact that the ANOVA evaluated the arcsine-transformed data of the individual listeners, while Fig. 2 shows the mean data pooled across all the listeners.

Arbogast, T., Mason, C., and Kidd, G. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Boersma, P. (**1993**). "Accurate short term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, 97–110.

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

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

Bronkhorst, A. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acustica **86**, 117–128.

Bronkhorst, A., and Plomp, R. (**1988**). "The effect of head-induced interaural time and level difference on speech intelligibility in noise," J. Acoust. Soc. Am. **83**, 1508–1516.

Brungart, D. (**2001a**). "Evaluation of speech intelligibility with the coordinate response measure," J. Acoust. Soc. Am. **109**, 2276–2279.

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

Brungart, D., and Simpson, B. (**2002**). "Within-ear and across-ear interference in a cocktail-party listening task," J. Acoust. Soc. Am. **112**, 2985–2995.

Brungart, D., Simpson, B., Ericson, M., and Scott, K. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Cherry, E. (**1953**). "Some experiments on the recognition of speech, with one and two ears," J. Acoust. Soc. Am. **25**, 975–979.

Conway, R., Cowan, N., and Bunting, M. (**2001**). "The cocktail party phenomenon revisited: The importance of working memory capacity," Psychon. Bull. Rev. **8**, 331–335.

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

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

de Cheveigne, A. (**1993**). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. **93**, 3271–3290.

Dorman, M., Loizou, P., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Drullman, R., and Bronkhorst, A. (**2000**). "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," J. Acoust. Soc. Am. **107**, 2224–2235.

Ebata, M. (**2003**). "Spatial unmasking and attention related to the cocktail party problem," Acous. Sci. Technol. **24**, 208–219.

Egan, J., Carterette, E., and Thwing, E. (**1954**). "Factors affecting multichannel listening," J. Acoust. Soc. Am. **26**, 774–782.

Ellis, D. (**2003**). "Sinewave Speech Analysis/Synthesis in MATLAB," http://www.ee.columbia.edu/dpwe/resources/matlab/sws/

French, N., and Steinberg, J. (1947). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Freyman, R., Balakrishnan, U., and Helfer, K. (2001). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Freyman, R., Helfer, K., McCall, D., and Clifton, R. (1999). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3587.

Hawley, M., Litovsky, R., and Culling, J. (2004). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.

Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Kidd, G. J., Mason, C., Rohtla, T., and Deliwala, P. (1998). "Release from informational masking due to the spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **104**, 422–431.

Kidd, G. J., Mason, C., Arbogast, T., Brungart, D., and Simpson, B. (2003). "Informational masking caused by contralateral stimulation," J. Acoust. Soc. Am. **113**, 1594–1603.

Moore, T. (1981). "Voice communication jamming research," in AGARD Conference Proceedings 331: Aural Communication in Aviation, pp. 2:1–2:6. Neuilly-Sur-Seine, France.

Moray, N. (1959). "Attention in dichotic listening: Affective cues and the influence of instructions," Q. J. Exp. Psychol. **9**, 56–60.

Neff, D. (1995). "Signal properties that reduce masking by simultaneous random-frequency maskers," J. Acoust. Soc. Am. **98**, 1909–1920.

Peissig, J., and Kollmeier, B. (1997). "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," J. Acoust. Soc. Am. **35**, 1660–1670.

Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (1981). "Speech perception without traditional speech cues," Science **212**, 947–950.

Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Stern, R. (1998). "Robust speech recognition," in *Survey of the State of the Art in Human Language Technology*, edited by R. Cole, J. Mariani, H. Uszkoreit, G. Varile, A. Zaemen, A. Zampolli, and V. Zue (Cambridge University Press, Cambridge).

Treisman, A. (1960). "Contextual cues in selective listening," Q. J. Exp. Psychol. **12**, 242–248.

Tyler, R., Preece, J., and Tye-Murray, N. (1986). "The Iowa audiovisual speech perception laser videodisk," in *Laser Videsodisc and Laboratory Report* (Department of Otolaryngology, Head and Neck Surgery, University of Iowa Hospital and Clinics, Iowa City, IA).

Wightman, F., Callahan, M., Lutfi, R., Kistler, D., and Oh, E. (2003). "Children's detection of pure-tone signals: Informational masking with contralateral maskers," J. Acoust. Soc. Am. **113**, 3297–3305.

Wood, N., and Cowan, N. (1995). "The cocktail party phenomenon revisited: Attention and memory in the classic selective listening procedure of Cherry," J. Exp. Psychol. **124**, 243–262.

Yost, W. (1997). "The cocktail party problem: Forty years later," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Erlbaum, Hillsdale, N.J.), pp. 329–348.

Zurek, P. M. (1993). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., edited by G. Studebaker and I. Hochberg (Allyn and Bacon, Portland).