How does a listener resolve linguistic properties conveyed by speech? Many descriptions attribute a role to brief spectral details in narrow frequency ranges. Perceptual standards allow far more variety, revealed by the robustness of perception of distorted speech. The present study considered effects of spectral blur on the recognition of words. Listeners heard presentations of noise-vocoded Easy and Hard words. The number of spectral channels increased with each presentation, reducing blur within a trial. Test conditions varied the severity of initial blur, either 1 or 5 channels. In all conditions, the final presentation had 9 bands, yielding a net blur reduction of 4 or 8 bands. A control used two repetitions of each word at 9 spectral bands. Across the tests, exposure to spectral blur impaired the recognition of Easy and Hard words alike regardless of the listener's initial exposure. The pattern of results exposes the role of attention, uncertainty and spectral resolution in the phonetic contribution to word identification.
PERCEPTUAL ROBUSTNESS

A spoken message can be understood under conditions that differ greatly from careful speech spoken in a quiet room to an attentive listener. Whether the compromise is caused by a talker’s casual articulation, a noisy enclosure, or a listener’s hampered acuity, the robustness of utterances in conveying linguistic and personal attributes tempers our technical descriptions and explanations of speech perception. Studies of sine-wave speech (Remez, Rubin, Pisoni & Carrell, 1981), chimerical speech (Smith, Delgutte & Oxenham, 2002) and noise-band vocoded speech (Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995) have revealed wide perceptual tolerance of the absence of the natural acoustic correlates of phoneme contrasts when the coarse-grain spectrotemporal pattern of speech is preserved. The adequacy of such elaborately distorted signals depends on two independent aspects of the perception of utterances. First, the time-varying properties of speech in the absence of typical short-term characteristics can be sufficient to initiate a perceptual projection from sensory attributes to phonetic attributes. Second, the nested constituents of linguistic organization, from syllable to word to phrase, remain available to govern a perceiver’s attention whether the sensory form of speech is typical or anomalous. A speech wave that has been distorted might leave a perceiver uncertain about the specific sound that conveys a familiar segmental contrast. However, at any moment in an evolving linguistic impression of speech, the sonorance of the syllable structure is ordinary, the composition of words expresses the regularities of the language, and the short-term procession of words in phrases is often patterned grammatically. These several sources of knowledge permit adaptation to variation in the short-term properties of sensory samples that express the spoken contrasts (see Pardo & Remez, 2005).

The study reported here aimed to develop an assay of the convergence of sensory and cognitive functions in speech perception, examining the specific instance of lexical identification. Under ordinary circumstances, the recognition of a word obliges a perceiver to resolve a familiar item, a word, from an unfamiliar sensory form. That is to say, the phoneme series that expresses the word CHECK, to take an example, is linguistically determined and universal among English speakers. Nonetheless, no talker speaks phonemes. Instead, speech production consists of a phonetic realization of a phonemic abstraction specific to a talker’s size, age, vitality, social aims, regional dialect and mood, among other attributes, and this determines the acoustic properties available for perception (Remez, 2010). To develop the measure of sensory and cognitive convergence, a classic finding from the vision lab was adopted as a model. In addition, a recent innovation in spoken word recognition studies was used to sharpen the empirical analysis of the role of lexical processes in perception.

Combining Two Assays

Blur Tolerance in Recognition

A classic project of cognitive psychology (Bruner & Potter, 1964) had examined the identification of objects depicted in photographs, and served as a model for the present investigation. The procedure varied the amount of optical blur in an image, ranging from great initial blur precluding the recognition of objects to more moderate blur permitting identification. In a trial, a blurry initial image was presented and a subject was asked to identify the depiction. Continuing over several trials, the image was presented with diminishing blur, with the observer attempting to identify the depiction as the image became sharper. The effect of blur on identification was systematic: the distinctness of visual contours grossly predicted identification performance, but there was an intriguing qualification. In a test block, some images were first presented at extreme blur, while others were first shown at more moderate blur. The greater the duration or the extent of initial blur in an image, the sharper the focus that was required to identify the depiction. In explaining the difference in blur tolerance as a consequence of immediate experience, the authors appealed to a cognitive function, proposing that recognition inherently includes the use of visual analysis to confirm or to reject the hypothetical identity of a depicted object. Put simply, the finding shows that greater initial blur impedes the rejection of false impressions of depiction, permitting such errors to persist at lesser blur. The corollary claim about perception is that hypotheses about real objects and events guide analysis of a visible scene; the poorer the resolution of sensory detail, the harder it is to reject a false hypothesis.

Despite occasional analogies in the perceptual literature between visual and auditory scene analysis, including the constituent function of object recognition, there have been few attempts to create a test of the perception of utterances analogous to this classic of visual perception. The test reported here employed a noise-band vocoder (Tiger Speech Technology) to represent the spectrum of spoken words in different resolution over 5 kHz, varying from 1 band (extreme blur precluding word recognition) to 9 bands (moderate blur sufficient for word
FIGURE 1. Effects of noise-band vocoder: Spectrograms of (A.) natural utterance of the word CHECK; (B.) 18-band vocoded version; (C.) 9-band vocoded version; (D.) 1-band vocoded version.
identification). Figure 1 shows the spectrogram of a natural utterance of a spoken word in panel A: the neighboring panels show three variants differing in the number of analysis bands and, therefore, the varying spectral blur imposed on the vocoded versions. Had our test of word identification included this manipulation alone, it would have offered a straightforward analogy to the original study of visual object identification on which it was modeled. However, a flanking method was used in an effort to distinguish overall effects of cognitive fluency from the acuity of perceptual analysis in the recognition of words. This additional method used the comparison of Easy and Hard word sets.

*Recognizing Easy and Hard Words*

Easy and Hard word sets were originally developed (Bradlow & Pisoni, 1999) to provide a calibrated measure of signal-independent influences on spoken word recognition. In contrast to spectral blur, an acoustic property of words, the contrast between Easy and Hard words exploits the inhomogeneity of the English lexicon, independent of the characteristics of an acoustic signal or its sensory samples. A word in the Easy set is more common according to an estimate of its frequency of occurrence; it is similar to relatively few other words; and, all of the words to which it is similar are less frequent than the word itself. In contrast, a Hard word is less frequent, is similar to relatively many words, and the words to which it is similar are all more frequent than the word itself. All other things being equal, identifying a Hard word is a subtler challenge than identifying an Easy word because it must be distinguished from a greater number and variety of other known words. Indeed, with matched signal characteristics, performance level is lower for Hard words than Easy. For the present purposes, identification of Easy words can be taken as an index of cognitive fluency, in contrast to Hard words, which offers an index of the acuity of perceptual analysis. By combining the manipulation of spectral blur and the signal-independent requirements for distinguishing a known word from lexical neighbors, this test sought to assay the contributions of perceptual analysis and cognitive function in spoken word recognition.

*The Empirical Project*

*Spoken and Vocoded Test Materials*

Natural speech samples of 74 Easy and 74 Hard words, produced in New York dialect by an adult male, were used to compose the test. These were sampled to disk, analyzed, and vocoded using AngelSim CIS (Tiger Speech Technology), software for analysis and synthesis that permitted a variable number of bands to be set, all excited by noise. Six variants were created of each word: 18-bands, approximating the natural spectral resolution of the human auditory system; and, in order of decreasing blur, 1-3-5-7- and 9-bands. The bands were linear and contiguous across frequency, and in every case a noise source was used to excite the bands in proportion to the instantaneous energy within the band limits of the natural sample.

*The Recognition Tests*

Several tests were conducted. A control test used 18-band vocoded versions of 74 Easy and 74 Hard words. This test aimed to determine whether the mere conversion of natural samples to vocoded test items conserved the phonetic properties requisite for identification. A second test with different listeners used 9-band vocoded words, and aimed to estimate the identification performance with Easy and Hard words; in this test there was no preliminary exposure to versions of the words made unidentifiable by imposition of greater spectral blur. In a third test with a different group of listeners, a word was presented on each trial in a sequence of decreasing blur variants. A given word occurred first at extreme blur (1-band) or moderate blur (5-bands), and became progressively sharper in several steps, ending in a 9-band version. Listeners were informed that the sound of each word would become clearer over the course of a trial. At the end of each trial, a subject transcribed the word and a new trial began. In each test, the presentation order of the 148 words was random.

Listeners were drawn from the undergraduate population of Columbia University, and each participant received course credit in exchange for taking part in the procedure. None was familiar with vocoded speech or the protocols in the spoken word recognition. The tests were administered in groups of four listeners, who wore headphones in a sound-shielded room and who recorded the identity of the words by writing in a specially prepared booklet.
The Results

In the control test using 18-band vocoded versions, Easy and Hard words were identified with few errors. On this outcome, the mere use of the vocoder seemed to be innocuous, and testing proceeded to the next step in the project, to measure the effects of blur on identification. The method of the test compared two conditions: when subjects identified 9-band vocoded words with no prior exposure to instances of greater blur; and, 9-band vocoded words when instances of the words were initially presented at greater blur. In addition, two blur conditions, 1-band (extreme blur) and 5-band (moderate blur) were tested.

The group results of 9-band identification are presented in the first panel of Figure 2, with the dark bar portraying the average identification of Easy words and the light bar Hard words. Although performance with 9-bands declined relative to the error-free performance with 18-band items, performance remained strong, and by falling in the middle of the range there was opportunity to observe an effect of prior exposure to blur that facilitated or harmed performance. Easy words were identified better than Hard words, indicating that this degree of spectral blur was not adequate to promote the perceptual distinctiveness of Hard words within their lexical neighborhoods. Because individual subjects differed from each other in the words that were identified and those that were missed, the effect is best understood as a functional condition rather than as a loss of acoustic or sensory details that consistently blocks the identification of specific test items.

The test of identification after preliminary exposure to blurred items shows a main effect of blur exposure. Performance declined relative to the test of 9-band word recognition. Nonetheless, there was no effect of the extent of initial blur on ultimate identification, as shown in the right panel of Figure 2. Whether the initial blur experienced by listeners was extreme or moderate, performance was impaired equally, showing that any exposure to blur reduced
a listener’s ability to correct a false identification of the words. Because both Easy and Hard word recognition declined in step, this effect of initial blur does not seem to be due simply to cognitive load, which would have disadvantaged the Easy words; nor, to a simple loss of acuity in the auditory-to-phonetic projection, which presumably would have harmed performance especially with Hard words.

**WORD RECOGNITION AND VISUAL OBJECT IDENTIFICATION**

The premise of the test devised here is an analogy between recognizing spoken words and recognizing objects visually. While the component cognitive functions are parallel — indeed, much of the inspiration of functions and mechanisms in cognitive psychology derives from studies of speech and vision — the attempt for an analogy in experimental method failed to find a clear dose-response effect of acoustic blur that paralleled the effect in visual object recognition. It is not clear whether this is due to nuisance variables in empirical method, or to intrinsic differences in the perception of spoken words and the perception of objects by vision. Nonetheless, the suppressive effect of blur exposure on Easy and Hard word recognition alike indicates the operation of the function in spoken words described first for visual scenes nearly fifty years ago.

Certainly, before concluding that the difference between speech perception and visual object perception is due to an intrinsic difference between these two functions, the evidence should be sought from test techniques supplementary to those reported here. For one, it will be useful to compare an identification threshold measure, despite its demerits (Haber & Hershenson, 1965), to the procedure used here. This might determine whether observing a dose-response effect of blur is contingent on an identification threshold measure. For another alternative technique, a different portion of the potential range of spectral blur might be informative. The appearance of a dose-response of blur might be contingent upon overall performance level, which varies with the ultimate level of spectral blur. It is also conceivable that abbreviating the test would permit a differential effect of prior blur exposure to appear. If short-term learning occurs over the course of 148 test trials, this might obscure an effect early in the procedure that is contingent on the extent of initial spectral blur. Through the use of converging measures, this method promises to expose the interaction of sensory analysis of speech and the cognitive functions of lexical identification.

**ACKNOWLEDGMENTS**

The authors gratefully acknowledge the assistance of Aislinn Crank, Rebecca Giglio, Stavrourla Koinis and Andrea Wycoff. This research was supported by a grant from the National Institute on Deafness and Other Communication Disorders (DC000308). During the term of the project, C. B. C. was supported by an award from the Howard Hughes Biomedical Science Pipeline Project.

**REFERENCES**


