Effects of Stimulus Variability on Recall of Spoken Sentences:
A First Report

John R. Karl and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

1This research was supported by NIH Research Grant R01 DC-00111 and an NSF Graduate Student Research Fellowship. A shorter version of this paper was presented at the 127th meeting of the Acoustical Society of America, Boston, MA (May 1994).
Abstract

Previous research has shown that talker-specific information in the speech waveform affects many stages of processing—from perception and working memory to recall. In perception, trial-to-trial variability in the speech signal requires a perceptual adjustment that affects identification performance. This perceptual adjustment may be thought of as a form of normalization of variability in the signal. Whatever form this normalization takes, it apparently does not filter out talker-specific information to uncover abstract, invariant linguistic units; as shown in both perception and memory studies, detailed information about the source signal is retained in memory for spoken words. The present investigation was conducted to determine whether the observed effects of talker-specific information on the processing of spoken words and isolated syllables are also evident in the processing of fluent sentences. In three experiments, subjects listened to and transcribed lists of sentences, spoken either by a single talker or by many talkers. After the transcription task, subjects recalled the sentences in a probed recall paradigm. The probe cue contained either no talker-specific information (i.e., it was presented visually) or it contained talker-specific information (i.e., it was presented auditorily). The experiments varied slightly in terms of design and instructions given to the subjects. The results indicate that the effects of variability in spoken sentences on perception, memory, and other cognitive processes are similar to effects found in earlier studies with spoken words. Implications of these findings for symbol-based and non-analytic accounts of perception and memory are discussed.
Effects of Stimulus Variability on Recall of Spoken Sentences: A First Report

Introduction

The acoustic waveform that carries a given linguistic message is notoriously variable across different speakers and different utterances from a single speaker (Carrell, 1984; Fant, 1973; Ladefoged, 1980; Stevens & Blumstein, 1981; Summerfield & Haggard, 1973). Despite this high variability in the signal, the linguistic percepts of listeners remain constant for a given message. Explaining this perceptual constancy in the face of tremendous variability in the acoustic signal has been a major challenge for theories of speech perception (Klatt, 1986; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Pisoni, 1992; Remez, 1994; Shankweiler, Strange, & Verbrugge, 1976). In general, theorists have assumed that variability in the speech signal poses a problem for listeners: It has been assumed to impede perceptual identification of the linguistic units that are used in the syntactic and semantic processing of an utterance. Many theorists have widely concurred that the perceptual system must solve this problem by identifying invariant cues to the linguistic units that comprise the message (e.g., Blandon, Henton, & Pickering, 1984; Disner, 1980; Gerstman, 1968; Green, Kuhl, Meltzoff, & Stevens, 1991; Liberman et al., 1967; Shankweiler et al., 1976). The goal of early perceptual processing of speech has thus been assumed to be the conversion of the speech signal into a sequence of abstract, invariant linguistic symbols that are identical across all speakers and utterances. Such "normalizing" of the speech signal into abstract linguistic units has been a theoretical cornerstone of speech perception research for many years (Abercrombie, 1967; Joos, 1948; Laver & Trudgill, 1979; Shankweiler et al., 1976; Studdert-Kennedy, 1974; Verbrugge, Strange, Shankweiler, & Edman, 1976). A recent quote from Halle (1985) is illustrative:

...it is unlikely that the information about the phonic shape of words is stored in the memory of speakers in acoustic form resembling, for instance, an oscillogram or a sound spectrogram. One reason that this is improbable is that when we learn a new word we practically never remember most of the salient acoustic properties that must have been present in the signal that struck our ears; for example, we do not remember the voice quality of the person who taught us the word or the rate at which the word was pronounced. Not only voice quality, speed of utterance, and other properties directly linked to the unique circumstances surrounding every utterance are discarded in the course of learning a new word. (pp. 101-102)

Two important predictions follow from the assumption that the speech signal is converted into abstract, idealized linguistic symbols. First, the representation of speech in memory will be independent of its acoustic source. This assumption entails that the representation of a given utterance will be equivalent across different acoustic manifestations. Second, the perceptual processes subserving normalization will operate independently of the amount of experience a listener has with a particular voice. That is, the processes involved in converting an acoustic signal into idealized units will not be affected by exposure to a particular voice. Throughout the research literature, however, one finds reports suggesting that these predictions are not supported by experimental evidence. One line of research indicates that speech signals are represented in the nervous system in a manner that preserves many fine perceptual details of the original acoustic signal (Goldinger, 1992; Goldinger, Pisoni, & Logan, 1991; Lightfoot, 1989; Martin, Mullennix, Pisoni, & Summers, 1989; Nygaard, Sommers, & Pisoni, 1992, 1994; Palmeri, Goldinger, & Pisoni, 1993). Another line of research provides evidence that the on-line perceptual processing of speech is affected by a listener's previous exposure to a particular voice or utterance (Creelman, 1957; Mullennix & Pisoni, 1990; Mullennix, Pisoni & Martin, 1989; Peters, 1955; Sommers, Nygaard & Pisoni, 1992). Although the body of data that support this view is rapidly growing, much of it is based on experiments
employing small units of speech such as isolated words or syllables. Very little research has investigated
issues related to the role that talker-specific or "indexical" information plays in the processing and the
memory representation of spoken sentences. This report describes recent research on the contribution
talker-specific information to processing spoken sentences and its relevance to recent non-analytic or non-
symbolic models of perception and memory (e.g., Brooks, 1978; Jacoby & Brooks, 1984; Kolers &

To place the present study in context, we will briefly examine the theoretical framework that has
guided research and theories in speech perception for nearly 50 years. This general framework has stemmed
from abstractionist or analytical assumptions about cognition. The abstractionist view holds that mental
content consists of abstract symbols which "contain" meaning and are manipulated according to prescribed
rules. These symbols are taken to be abstract tokens that contain meaning by referring to external events or
symbolic view within speech perception has received support from three major sources: (a) the influence of
formal linguistics; (b) the problem of variability in the speech waveform; and (c) the symbolic view of
cognition that has dominated most other areas of cognitive psychology (Fodor, 1983; Gardner, 1987;
Haugeland, 1981, 1985; Neisser, 1967; Posner, 1969; Pylyshyn, 1984). We will look at each of these
influences in turn.

Researchers in speech perception have naturally drawn ideas about the components and structure
of language from formal linguistics. Linguistics has shown language to consist of a complex but patterned
system of units at many levels—from low-level units such as phonetic features up through phones,
phonemes, morphemes, words, phrases, and sentences. Language is viewed as a system comprised of these
abstract symbols with accompanying rules for manipulating them. It has sometimes been thought that
elucidating the structure, regularities, and properties of this system will naturally lead to knowledge about
the structures and processes of human cognition (Chomsky, 1972; Fodor, 1983). Frequently, however,
language has been thought of as a complex, interesting system to be studied as an end in itself. In either
case, the dominant approach has been formalist, viewing language as an abstract system that can be studied
independently of questions about how the human nervous system actually perceives, comprehends, or
produces language. In Chomsky's well-known terminology, linguistics has taken as its object of study the
"competence" of language users rather than their "performance" (Chomsky, 1965). The system of symbols
that is described by this approach is thought to be a separate entity from the acoustic medium that happens
to carry those symbols. Thus, a distinction is drawn between linguistic and non-linguistic aspects of the
speech waveform. Linguistic attributes are considered to be those features of the waveform that specify
linguistic units, whether features, phones, phonemes, or other units. In this view, the speech signal also
carries non-linguistic information as well—information reflecting the gender, age, emotional state, or
dialect community of the speaker. This information is called "indexical" (Abercrombie, 1967) or "para-
linguistic" and is thought of as incidental to the linguistic content of an utterance (e.g., Halle, 1985).

These theoretical stances have heavily influenced psychologists interested in how humans perceive,
comprehend, remember, or produce language. The cognitive operations subserving spoken language have
been assumed to take as their units of processing one of the abstract, idealized linguistic units proposed by
formal linguistics (e.g., Liberman, et al., 1967; Peterson, 1961; Studdert-Kennedy, 1974; Klatt, 1986).
Indexical information has been thought to be irrelevant to the cognitive processing of language, or to be
non-linguistic in nature (Abercrombie, 1967).

The second influence favoring abstractionist assumptions in speech perception has arisen from
within the field of speech perception itself—the problem of the great variability of the speech waveform.
When the sound spectrograph first made a visual representation of the speech waveform possible, researchers expected to quickly identify those aspects of the waveform that had linguistic significance (Liberman, 1957; Liberman et al., 1967). However, they soon discovered that there are no obvious invariant acoustic correlates to the phonetic value of speech sounds (Goldinger, et al., 1995; Liberman et al., 1967; Pisoni, 1993; Stevens & Blumstein, 1981). Rather, the speech waveform displays great variability in its mapping to discrete linguistic categories (Klatt, 1986; Pisoni, 1993; Stevens, 1972). Many sources contribute to this acoustic variability both within and between speakers, including the size and shape of an individual's vocal tract, differences in articulatory strategies used to realize speech sounds (Ladefoged, 1980; Nygaard et al., 1994; Stevens, 1972; Stevens & Blumstein, 1981), and differences in glottal function (Carr & Trill, 1964; Carrel, 1984; Klatt, 1986). Researchers have described in detail the enormous acoustic variability found in vowels (e.g., Fant, 1973; Joos, 1948; Peterson & Barney, 1952; Summerfield & Haggard, 1973) and in consonants (e.g., Fourcin, 1968; Rand, 1971). Despite this variability, however, listeners have little trouble uncovering the linguistic units of speech encoded in the acoustic waveform. Listeners are, in fact, seldom aware that any perceptual processing is necessary.

To account for this perceptual constancy despite severe acoustic variability, numerous theorists have proposed sensory and cognitive mechanisms that compensate for variability by some process of perceptual normalization. Although the precise details of this process are usually unspecified, several preliminary proposals have been made. Normalization is sometimes conceived of as an early perceptual process that filters out the talker-specific attributes of the speech signal, converting the signal into invariant linguistic units that are then passed on to higher levels of analysis (Halle, 1985; Shankweiler et al., 1976; Studdert-Kennedy, 1974, 1976; Verbrugge, et al., 1976). After this early perceptual normalization, the remainder of cognitive processing is assumed to operate over abstract symbolic units. In other formulations of normalization, it is not clear whether it is assumed to operate early in processing to filter out talker-specific information, or instead whether it operates in parallel with processes that store speech in memory in a holistic fashion, retaining fine perceptual details of the acoustic waveform (e.g., Klatt, 1986; Nearey, 1989; Shankweiler et al., 1976; Strange, 1989a). Regardless of how normalization has been characterized, the focus of research in speech perception has been on finding the invariant cues in the acoustic waveform that are presumed to be mapped onto abstract, symbolic units. Although this approach has not met with a great deal of success, it continues in one form or another today (Kuhl, Williams, & Meltzoff, 1991; Laver & Trudgill, 1979; J. D. Miller, 1989; Repp, 1982; Stevens & Blumstein, 1981; Strange, 1989a, 1989b).

The third factor contributing to the prevalence of symbolic views of speech processing has been the easy fit of both formalist analyses of linguistics and the normalization processes proposed in most models of speech perception into the prevailing symbolic theories of cognition in areas outside of speech perception, primarily those related to vision and visual processing of language (e.g., Haber, 1969; Morton, 1970; Neisser, 1967; Norman, 1968; Norman & Rumelhart, 1970; Posner, 1969; Pylyshyn, 1984). Many of these proposals have relied extensively on direct or indirect analogies to symbolic computation as performed by computers and on computer implementations of cognitive models (Gardner, 1985; Haugeland, 1981, 1985).

The analytic or symbolic view of speech perception has had significant consequences for the experimental methodologies that have been employed in research. Because variability has been thought of as a source of noise that obscures the linguistic content of the signal, experimenters have generally sought to minimize this variability by employing as few speakers or tokens as possible for use as stimuli in perceptual experiments. Consequently, the effects of the natural variability in the speech signal have been kept to a minimum in the laboratory. When effects were observed, they were considered to be due to the noise of the signal and to be irrelevant to the interpretation of data (Elman & McClelland, 1986). Thus,
researchers have often avoided confronting the effects of variability and their implications for models of speech processing.

Throughout the years, however, researchers have regularly reported the effects of variability on both the perceptual processing of speech and on the representation of speech in memory. The perceptual consequences of variability were first observed in 1955, when R. W. Peters (1955) reported that, across a range of signal-to-noise ratios, lists of messages produced by multiple talkers were consistently less intelligible than the same lists produced by single talkers. Several years later, C.D. Creelman (1957) published results of a study in which he found that, at several signal-to-noise ratios, articulation scores (i.e., perceptual identification scores) were worse for lists of phonetically-balanced words produced by multiple talkers than they were for the same lists produced by single talkers. Neither of these reports received much attention, and their authors did not pursue the questions they raised. Several years later, however, a series of reports was published that began to establish the significance of variability for perceptual processing of speech. Summerfield & Haggard (1973) found that vowels synthesized from a single talker were categorized faster than those synthesized from multiple talkers. Verbrugge et al. (1976) found that natural vowels were identified more accurately when they were all produced by a single talker than when they were produced by multiple talkers. The results of these perceptual studies provide support for the proposal that talker-specific information in the speech signal may play an important role in the on-line processing of speech. Generally, adjusting to different voices from trial to trial within an experiment appears to require perceptual resources that are not needed when the voice remains constant (Mullennix et al., 1989). This is shown by slower reaction times or less accurate identification of stimuli when the voice changes from trial to trial. Despite several demonstrations of such effects in the literature, however, there has been little discussion of what the nature of the associated perceptual processes might be (see however Pisoni, 1993).

Another line of research has provided evidence that the long-term memory representation of spoken words incorporates details of the talker-specific or indexical information in the original acoustic signal. This has typically been demonstrated by showing that, after studying a list of words, subjects subsequently recognize, recall, name, or identify the same words more accurately or faster when they are presented in the same voice as the original stimuli, rather than a different voice. For example, Tulving & Colotla (1970) examined free recall for lists of words that were spoken in either a single language or in three different languages. They found that early list items from the single-language lists were recalled more accurately than those from the three-language lists. In 1974, Craik and Kirsner employed a same/different matching task to show that subjects retain talker-specific information in long-term memory for up to two minutes. They found that listeners were faster at deciding whether a stimulus word had been previously presented (i.e., was the "same" as a previous word) when test words were presented in the same voice as in the original presentation. Furthermore, subjects could explicitly judge whether the voice of the stimuli was "old" or "new." Cole, Coltheart, and Allard (1974) and Allard and Henderson (1975) reported similar results.

Several studies have focused on memory for the input mode of stimuli, and in general, the results show that stimuli are stored in memory with an indication of the mode in which they had originally been presented. Bray & Batchelder (1972) found that, although the mode of presentation (auditory or visual) had no effect on the accuracy of subjects' recall of nouns, subjects nevertheless recalled the mode of presentation with better than chance accuracy. Fisher & Cuervo (1983) examined subjects' memory for indexical attributes—or "physical features" in their terms—of spoken sentences as a function of the relevance of those attributes to the comprehension of the sentences. They found that the gender of a speaker
or the language of a sentence was remembered better when those features were instrumental to the comprehension of the sentence.

Geiselman & Bellezza (1976, 1977) found that subjects were able to judge whether repeated sentences had originally been spoken in male or female voices. Their conclusion was that the gender of a speaker connotes information about the meaning of a sentence, and is thus stored in memory as part of the semantic interpretation of a sentence. In this proposal, voice information becomes encoded in memory as an abstract, semantic designation of the gender of a speaker.

Recently, the theoretical significance of these findings has been pursued in a series of experiments conducted in our laboratory. These studies have examined both the long-term memory representation of spoken words and the on-line perceptual effects of talker variability. In one study of talker variability, Mullennix et al. (1989) replicated the findings of Peters (1955) and Creelman (1957) described above. In a perceptual identification experiment, talker variability and lexical density were manipulated. Across three speech-to-noise ratios, using words from both high and low density neighborhoods, items from multiple talkers were identified less accurately than items from single talkers. In a second experiment, Mullennix et al. extended Peters' and Creelman's work by examining the effects of talker variability on the time course of perceptual processing. Subjects heard words over headphones, and were required to repeat them out loud as quickly as they could. The latency of naming words from high and low density neighborhoods was measured in single- and multiple-talker conditions. Mullennix et al. found that responses were slowest in the multiple-talker, high-density condition, again providing evidence that these sources of variability affect on-line perceptual analysis of the input, and, specifically, suggesting that some form of perceptual adjustment or "attunement" is required when switching from one voice to another.

Other experiments in our laboratory have investigated the effects of talker variability on memory for spoken words. Martin et al. (1989) found that remembering multiple-voice lists of words requires more resources in working memory than do single-voice lists. Multiple-voice lists decrease the speed and/or efficiency of rehearsal processes involved in the transfer of spoken words from working memory to long-term memory. In their first experiment, Martin et al. used a 10-item list of spoken words in a serial recall task. The lists were spoken by either a single talker or by multiple talkers. Recall of items was better in the primacy portion of the lists in the single-talker condition than in the multiple-talker condition. (Items from the middle and end of the lists were assumed to be recalled from working memory rather than from long-term memory; see Rundus, 1971). Thus, rehearsal processes in working memory appeared to be less efficient in the multiple-talker condition than in the single-talker condition. To further investigate this effect, a second experiment manipulated the demands on working memory. Subjects were presented with a memory load of zero, three, or six digits before the presentation of the 10 spoken words, and were instructed to remember the digits throughout the presentation of the word lists. After the words were presented, subjects attempted to recall both the digits and the words. Recall for the digits was better for single-talker lists than for multiple-talker lists, supporting the idea that multiple-voice lists require more resources for rehearsal than do single-voice lists. As in the first experiment, recall for words in the early portion of the lists was better in the single-talker condition than in the multiple-talker condition. In their third experiment, Martin et al. determined that the voice effect in recall of spoken words was not due to subjects' use of talker-specific voice information as a search or retrieval cue during recall. Subjects were given an arithmetic task to complete after the presentation of the 10-item lists. This ensured that talker-specific voice cues could not remain in working memory at the time of recall. This manipulation had no effect on the difference between recall in the single- and multiple-talker conditions; recall was again best for the single-talker in the primacy portion of the lists. Thus, this difference was not due to the use of talker-specific voice information as search or retrieval cues. Rather, it was due to the cost of multiple voices to
processes involved in the transfer of items from working memory to long-term memory. This effect has been replicated in a number of other studies (Lightfoot, 1989; Logan & Pisoni, 1987).

As Martin et al. (1989) explained, their study did not provide conclusive evidence regarding the locus of the talker variability effect on serial recall of words; their results were consistent with effects of talker variability operating during early perceptual encoding, during rehearsal in working memory, or both. To further examine the locus of the effect of talker variability, Goldinger et al. (1991) replicated Martin et al.'s experiments and extended them with two new variables. The first new variable was the phonetic confusability of the words within a list. This manipulation was known to operate during early perceptual encoding of spoken words (Luce, 1986; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985). The second new variable was the rate of presentation of the words within a list. Rate of presentation interferes with rehearsal processes in working memory (Glanzer & Cunitz, 1966; Jahnke, 1968; Murdock, 1962; Rundus, 1971). Because varying the presentation rate of the words affects the rehearsal processes for early list items, Goldinger et al. predicted that slow presentation rates would allow more elaboration than fast rates. If talker variability produces its effect during rehearsal, one would expect to find an interaction between the variables of presentation rate and talker. This is, in fact, what Goldinger et al. found. The change in presentation rate affected recall of words from multiple-talker lists more than words from single-talker lists. At the fast rate used by Martin et al. (1989), single-talker lists were recalled better than multiple-talker lists. However, at slower rates the reverse was found: multiple-talker lists were recalled better than single-talker lists. The word-confusability variable was assumed to affect only early perceptual encoding of the stimulus words, and consequently no interaction with presentation rate was expected. The results were consistent with this prediction as well. This pattern of results indicated that, compared to single-voice word lists, multiple-voice word lists cause deficits in both the perceptual encoding of spoken words and the rehearsal processes necessary for transferring items to long-term memory.

In another series of experiments, voice information and phonemic information were found to be processed as integral components of each other (Mullennix and Pisoni, 1990). Performing a Garner speeded-classification task (Garner, 1973), subjects were required to selectively attend either to the voice of a stimulus or to its phonemic identity while the non-attended dimension varied randomly. Whether attending to the voice or to the phonemic dimension of the stimuli, subjects were unable to ignore the variation in the non-attended dimension: voices interfered with phoneme perception and phonemes interfered with voice perception. These results were interpreted as evidence that the voice dimension and the linguistic dimension of acoustic signals are not perceived independently, as assumed by symbolic accounts of perception (e.g., Abercrombie, 1967; Halle, 1985). Instead, voice information appears to be an integral component of the linguistic content of the signal.

The studies on the effects of talker variability on perception and on memory described above suggest that these components of processing are not independent but interact with each other. Several recent studies from our laboratory have investigated the nature of this interaction. Palmeri, Goldinger, & Pisoni (1993) employed a continuous recognition memory paradigm to investigate the effects of stimulus variability on perception and recognition memory. Palmeri et al. varied the voice in which words were repeated and the lag (number of items) between the original and repeated presentations of words. The voice of a repeated word was either the same as the original presentation, a different voice but the same gender, or a different voice and a different gender. Subjects were able to recognize repeated words faster and more accurately in same-voice repetitions across many different lag times; they were also able to explicitly judge whether words were new presentations, repeated presentations in the same voice, or repeated presentations in a different voice from the original presentation. These findings established several things. First, voice information is encoded automatically, not as a result of conscious or strategic effort. Second, voice
information is retained in memory for fairly long periods of time and affects perception of spoken words. And third, the representation of voice information in memory is not a generic code for gender but appears to preserve fine perceptual details of original stimuli.

More recently, Nygaard, Sommers, & Pisoni (1994) investigated the relationship between listeners' ability to identify speakers' voices and their ability to process speech by those speakers. After Nygaard et al. trained subjects to identify 10 talkers by name, they tested the subjects' perception of novel words. The experimental subjects were tested on novel words spoken by the same talkers who were used during training, whereas control subjects were tested on the same words spoken by novel talkers whose voices had not been used during training. Subjects in the experimental group were significantly better at perceptual identification of the novel words than were those in the control group. Thus, Nygaard et al. demonstrated that, in the course of becoming familiar with talkers' voices, listeners learn details about those voices that enhance their ability to perceive isolated words spoken by those voices. As Nygaard et al. put it, speech perception is a "talker-contingent" process.

With the exception Fisher and Cuervo (1983) and of the obscure report by Peters (1955), all of the above studies have been conducted using isolated words or syllables. It is not known whether the effects of talker variability would be observed on the more natural linguistic unit of sentences. Spoken sentences differ in many important respects from isolated words or syllables. These differences may affect perceptual processing at all levels, from acoustic-phonetic to syntactic and semantic levels. A major source of these differences is the fact that the acoustic waveform is often continuous across lexical boundaries in sentences (Klatt, 1986; Goldinger, Pisoni, and Luce, 1995). This continuity presents listeners with additional perceptual problems beyond those posed by isolated words. One additional problem is that listeners must identify the boundaries between words. Isolated words present no such problem, because the onsets and offsets of words are unambiguously defined by the silence that precedes and follows each word. In fluent sentences, however, such clear acoustic correlates to lexical boundaries rarely exist. In addition to being acoustically continuous, the boundaries between words are also affected by the same factors that make acoustic phonetic processing problematical within isolated words—the lack of acoustic-phonetic invariance in the speech signal and the lack of linearity in segmental features (Pisoni & Luce, 1987). Lexical segmentation in sentences, then, poses additional perceptual problems beyond those encountered in isolated words or syllables.

At the same time, however, it appears that the cross-lexical continuity of the acoustic waveform in sentences may be of some perceptual benefit as well. Researchers have suggested that patterns of intonation across words may contain information that is useful for processing at many levels. For example, it may contain cues that are useful for phonetic-level processing (Haggard, Ambler, & Callow, 1970), as well as information that is helpful for defining word, phrase, and sentence structure (O'Shaughnessy, 1979; O'Shaughnessy & Allen, 1985). Suprasegmental information carried by the continuous waveform may even contain information that helps overcome the problem of lexical segmentation. Many researchers have suggested that perceptual mechanisms use suprasegmental information in segmenting the speech stream, both within words and across word boundaries (Geers, 1978; Lehiste, Olive, & Streeter, 1976; Nakatani & Schaffter, 1978; Norris, McQueen, & Cutler, 1994; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Scott, 1982; Streeter, 1978).

Sentences are also different from isolated words in the complexity of the linguistic information they contain. Sentences encode semantic and syntactic information that play important roles in perceptual analysis and comprehension. This high-level information appears to interact with lower-level acoustic-phonetic processes, both during perception and production. For example, talkers appear to continually
adjust their speech, balancing information provided by the segmental structure against that provided globally by the context of the sentence (Hunnicutt, 1985). Listeners, in turn, appear to use the semantic and syntactic structure of sentences to constrain their phonetic interpretation of the speech signal (Bard, Shillecock, & Alman, 1988; Cole & Jakimik, 1978; Grosjean, 1980; Grosjean & Gee, 1987; Lieberman, 1963; Tyler & Wessels, 1983). The syntactic structure of sentences also directly interacts with the acoustic realization of segments and words. For example, vowels spoken in sentences are approximately half the duration of vowels in the same words spoken in isolation (Pisoni & Luce, 1987). Factors influencing the durations of vowels in sentences include the locations of syntactic boundaries, syllabic stress, and the component features of adjacent segments (Gaitenby, 1965; Klatt, 1975, 1976; Lehiste, 1970). The importance of these acoustic and linguistic differences between words in sentences and words in isolation is illustrated by a study conducted by Salasso & Pisoni (1985). They found that words are typically identified correctly only 50% of the time when they are edited out of sentences and presented in isolation to listeners (see also Pollack & Pickett, 1963, 1964).

Sentences thus pose unique questions about the contribution of talker-specific information to spoken language processing. The present set of experiments address several of these issues. The first experiment was designed with two purposes in mind. First, we wanted to determine whether talker-specific information plays a role in the perception of sentences that is similar to the role it plays in the perception of isolated words (Creelman, 1957; Mullenix & Pisoni, 1990; Mullenix et al., 1989; Sommers et al., 1992; Nygaard et al., 1992, 1994). This will provide an opportunity to replicate the findings of Peters (1955), who showed that lists of sentences produced by multiple talkers were less intelligible than lists produced by single talkers. Such a result would suggest that sentence comprehension, like isolated word recognition, is a talker-contingent process that incurs costs to perceptual mechanisms as voice changes (Nygaard et al., 1992, 1994). The second purpose was to determine whether talker-specific information is encoded into the neural representation of sentences in a manner similar to the way it is encoded into the representation of isolated words (Craik & Kirsner, 1974; Martin et al., 1989; Goldinger, 1992; Goldinger et al., 1991; Lively, 1994). If the representation of sentences in memory does consist of a record of perceptual details, this would provide further evidence that analytic-symbolic accounts of speech perception and memory representation need to be reconsidered.

Three experiments were conducted to investigate these questions. Because the first two were very similar, we will refer to the experiments as Experiment I-A, Experiment I-B, and Experiment II. Each experiment employed a cued-recall paradigm that has been widely used in studies of memory for words (Roediger & McDermott, 1993). This paradigm begins with a study phase in which subjects are exposed to stimuli. Then, in a test phase, subjects receive probe cues for recalling the stimuli.

In the study phase of our experiments, subjects listened to a list of sentences and transcribed each one as they heard it. This task exposed the subjects to the stimulus sentences and provided a means of checking that they encoded each sentence correctly. In the test phase, subjects were presented with a probe cue—the first one or two words of each sentence—to help them recall the sentence. After receiving the probe cue, subjects were required to write down as much of each sentence as they could remember. In Experiment I-A, this recall task was a surprise to the subjects. In Experiment I-B, subjects were informed of the recall task at the outset of the experiment. The two experimental variables were the number of voices heard during study and the modality of the probe cue during the recall phase. The voice variable was either single-talker or multiple-talker. That is, during the study phase, subjects heard a list of sentences that were spoken either by a single talker or by many different talkers. The probe cue was either auditory or visual. In the auditory condition, subjects heard the cues to each sentence over headphones; in the visual condition,
subjects read the cues on a CRT display. Experiment II employed the same paradigm, with a slight change in experimental design.

Based on the earlier research described above, several possible outcomes were anticipated. If the effects of talker variability on the on-line processing of spoken sentences are similar to those that have been observed for spoken words, the talker variable should affect the accuracy of the transcriptions during the study phase. Specifically, transcription accuracy should be higher in the single-talker condition than in the multiple-talker condition. Also, if the effects of talker variability on the representation of spoken sentences in memory is similar to those that have been observed for spoken words, the information carried by the recall cues (lexical information alone in the visual cues, or lexical and voice information in the auditory cues) should affect the ability of subjects to recall the sentences. We expected that the auditory cues would elicit better recall performance than the visual cues because they contain both the linguistic information of the cue word and the indexical information associated with the particular talker who spoke the sentence. This additional indexical information might serve as a useful retrieval cue or retrieval context that could be used to match the cue to the sentence stored in memory (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981), or it could simply serve to restrict the search space in the multiple-talker condition.

EXPERIMENTS I-A AND I-B

Method

Subjects

Subjects were 160 undergraduate students enrolled at Indiana University who were paid $5.00 for their services. There were 58 females and 22 males in Experiment I-A, and 52 females and 28 males in Experiment I-B. None of the subjects reported any history of a speech or hearing impairment at the time of testing. All subjects were between the ages of 17 and 40 and all were native speakers of English.

Stimuli

The stimulus materials used in this experiment were taken from a digital database of 2,000 sentences that was created in our laboratory to study sources of variability in speech perception and spoken word recognition. The database consists of 10 male and 10 female talkers who recorded 100 Harvard sentences each (Egan, 1948). All materials consist of declarative sentences containing five key content words and several additional function words. The sentences are meaningful and syntactically normal. An example is "The yacht slid around the point into the bay." The key words in this sentence are underlined. The sentences are said to be phonetically balanced because the frequency of words and phonemes found in English is reproduced across successive lists (see Egan, 1948).

The talkers sat in a sound-attenuated IAC booth and spoke each sentence into a microphone as it was displayed on a computer monitor. They were instructed to read in a normal speaking voice. The order of the presentation of the 100 sentences was varied randomly for each talker. The analog signal was low-pass filtered at 10 kHz and digitally sampled at 20 kHz with 16-bit resolution using a DSC Model 240 analog-to-digital converter interfaced to a VAXstation 3500. After the digital recording, the stimulus files were edited using a digitally controlled waveform editor to remove the silence preceding and following the waveform. Overall RMS (root mean squared) amplitude levels for all sentences were digitally equated to ensure equal presentation levels.

After the database was complete, intelligibility scores for the 2,000 sentences were collected from a separate group of subjects. The 100 sentences from each of the 20 talkers were presented binaurally over
matched and calibrated TDH-39 headphones to 10 listeners at a comfortable listening level (75 dB SPL). The presentation was controlled by a PDP-11/34 computer. The listeners were told to transcribe each sentence as accurately as possible by typing it on a computer keyboard. Subjects were given as much time as needed to type each sentence. The resultant transcriptions were then scored for accuracy using only the five key words in each sentence. If one or more of the key words was incorrect, excluding obvious spelling or typographic errors, the entire sentence was scored as incorrect. Homonyms (such as their and there) were counted as correct. Any transcription that could be read as phonemically distinct from the target sentence was counted as incorrect (e.g., The place seems dull... was scored as incorrect for the target The play seems dull...). If all five key words were correctly transcribed, the sentence was counted as correct. Thus, each sentence could obtain a score of 0 to 10 for each talker, representing the number of listeners who transcribed that sentence correctly. With these scores, the average intelligibility of a given sentence as spoken by all 20 talkers was calculated, as well as the average intelligibility of a given talker across all 100 sentences.

For the present study, a subset of talkers and sentences was selected from this sentence database. Based on the results of the intelligibility study, the talker whose intelligibility was closest to the mean of the 20 talkers in the database was chosen for the single-talker condition. The sixteen most intelligible talkers of the remaining 19 were chosen for use in the multiple-talker condition.

Four sets of stimulus lists were then constructed: Lists A and B produced by a single talker and lists A and B produced by multiple talkers. Each list consisted of 48 sentences from the database. These sentences were chosen so that the mean and variance of the intelligibility scores would be equivalent for both the single-talker and multiple-talker test conditions. Across lists A and B, the mean score for the single-talker sentences was 8.96, and for the multiple-talker sentences it was 8.95. The standard deviation was 1.16 and 1.28 for single-talker and multiple-talker sentences, respectively.

Procedure

Upon arrival at the laboratory, subjects in Experiment I-A read a page of instructions explaining that the purpose of the experiment was simply to measure the intelligibility of the spoken sentences; the instructions for Experiment I-A did not mention the probed recall phase of the experiment that followed the transcription phase. In Experiment I-B, the instructions informed the subjects from the outset that there would be a recall test after the transcription. These instructions told subjects to transcribe each sentence as accurately as they could and then try to commit each one to memory. Three suggestions were given to the subjects in Experiment I-B to help them store the sentences in memory: form a mental picture of the scene described by the sentence, silently repeat the sentence several times, or listen to the voice of the speaker of each sentence. They were told that they would later receive part of each sentence as a probe cue to help them remember the sentences. In both parts of Experiment I, subjects were told to write out each sentence as accurately as they could. Subjects were told that the sentences were all normal English sentences consisting of common words that would be easy to understand. If subjects were not sure of what a speaker was saying, they were told to write their best guess if they had one, or to leave a blank for the word or words they could not understand. Subjects were told that there would be 48 sentences and that they would have 20 seconds to transcribe each one. They were told that correct spelling was not important, but that accuracy in transcribing the content of the sentences was very important. Subjects were also instructed to listen to the whole sentence before beginning to transcribe it.

During the experimental session, subjects heard a brief, 1000 Hz warning tone two seconds before each sentence was played over headphones. The twenty seconds given for the transcription of each sentence
was ample time for all subjects to write down their response. This fixed interval ensured that the duration of the transcription phase and the time between the transcription of a sentence and the probed recall of each sentence was equal for all subjects. The order in which the sentences were presented was randomized for each experimental session. There were 28 experimental sessions for each experiment, with one to five subjects in each session.

After the subjects transcribed the sentences, they were given the probed recall test. Additional instructions given at this time informed the subjects that a retrieval cue would be provided for each trial to help them recall the sentence. Subjects were told that the probe cues would be presented in the same order as the sentences had been during the transcription phase. In the Visual Cue conditions, the instructions explained that the probes would be presented on a CRT screen in front of the subjects. Subjects were told that the cue would consist of the first key word and any words prior to it in the original sentence. For example, in the sentence The boy was there when the sun rose, the first key word is boy. Thus, the words The boy... would be presented on the computer monitor in front of the subjects. If the first word of the sentence was a key word, then only that word was given as a probe cue. As in the transcription phase, a tone sounded two seconds before the cue appeared on the CRT. The visual cue itself was presented for two seconds so that it would be available to the subjects only briefly, as was necessarily the case with the auditory cues. In the Auditory Cue conditions, the probe cue for each sentence was presented binaurally over headphones. These cues were segmented out of the original sentence, so that the linguistic and indexical information of the cue was identical to the information in the original sentence. As in the Visual Cue condition, a tone sounded two seconds before each cue was presented. Twenty seconds were given for subjects to recall each sentence and write it out in the response booklet. Presentation of the study sentences and the probe cues were controlled by a PDP-11/34 computer.

Both the talker and the cue variable were manipulated using a between-subjects design. Subjects were randomly assigned to one of four conditions: (a) single-talker, visual cue; (b) single-talker, auditory cue; (c) multiple-talker, visual cue; and (d) multiple-talker, auditory cue. Within each of these conditions, subjects were randomly assigned to either List A or List B.

Results

Scoring

Sentences are highly complex stimuli which may be mentally represented at many levels such as surface features, syntactic structure, semantic content, etc. For this reason, several dependent measures were used to measure accuracy in the transcription and probed recall tasks. The first three scoring criteria we discuss were used in both Experiment I-A and Experiment I-B. The fourth was used only in Experiment I-B.

The strictest scoring criterion required every word in a sentence, including function words, to be transcribed or recalled exactly as it was in the stimulus sentence (apart from spelling errors and homophones). If any word was different from the stimulus, the sentence was scored as incorrect. This measured the exact linguistic form of a sentence, and will be referred to as the Exact Words scoring criterion. A second, less strict, criterion was designed to capture the most important linguistic content of the stimuli, but not every surface detail. Scoring by this criterion was based only on the five content (or key) words. Using this criterion, a sentence was scored as correct if all key words (five in the transcription task, four in the recall task) were transcribed or recalled correctly and they appeared in the same order as in the

---

2In a pilot study, it was found that typing the sentences into a computer proved too laborious for many subjects and distracted their attention from the task. Thus, subjects transcribed the sentences by hand in the present experiments.
original sentence. Function words were not evaluated. This will be referred to as the Key Words scoring criterion. The third basis for scoring was the most liberal. It was intended to capture the degree to which subjects transcribed or recalled the semantic content of the sentences, independent of their linguistic form. This scoring criterion will be called the Meaning criterion. This method necessarily required more judgment on the part of the scorer than the previous two methods. The criterion used for both the transcription and the recall data was as follows: a sentence was scored as correct if the scorer could form a mental picture from the sentence that corresponded in content to the original sentence. The primary semantic roles of agents, patients, and actions had to be represented in this picture for a sentence to be scored as correct. For example, given the stimulus sentence The swan dive was far short of perfect, any response that preserved the meaning of the sentence was counted as correct, no matter how different it was at the surface level. Thus, the following sentences would have been scored as correctly transcribed or recalled:

The swan dive was far from perfect.
The swan dive wasn't good.
The man's dive was bad.
The swan dive was not very graceful.

Responses that fundamentally changed the content of a mental picture of the scene were scored as incorrect. Examples of incorrect responses would be:

The swan dove far from perfectly.
The swan dive was far from shore.
The swan dive was perfect.

By this criterion, adjectives were generally deemed to be less important to the content of the sentence than nouns or verbs. Thus, adjectives might be altered or left out without fundamentally changing the content of a sentence, but nouns and verbs usually had to be either identical to those of the stimulus sentence or replaced with a synonym. They could not be left out of a response. This criterion differs from the other two in two three respects. First, it is based on the semantic content of the subjects' responses rather than the linguistic form, as are the other two criteria. Second, it is necessarily much less objective than the two form-based criteria. Third, it is an extremely lenient criterion, allowing considerable variation in subjects' responses. Although it may be a somewhat less reliable measure than the two form-based criteria, we included it in the experiment because we thought it might illuminate different aspects of subjects' performance than the other measures.

These three scoring methods were used in both the study and recall phases of each experiment. Because they all assign a score of one or zero to each sentence, the differences between subjects who remembered part of a sentence and subjects who remembered nothing of a sentence are obscured. Therefore, in Experiment I-B, a fourth criterion was used in order to capture something of the degree to which subjects accurately transcribed and remembered a given sentence. By this criterion, which we will call the Total Key Words measure, sentences were given a score corresponding to the number of key words that were transcribed or recalled correctly. Thus, in the transcription task, each sentence could receive a score from zero to five. A subject's score was then expressed as the percentage of total key words that were correctly transcribed across all sentences.

With all four scoring methods, scoring of the recall data was different in several respects from scoring of the transcription data. First, when scoring recall on the basis of the Key Words or the Total Key Words, the first key word was not counted because it was presented as the probe cue and was not recalled by the subjects. Thus, only the last four key words of each sentence were scored. In addition, the recall
phase was scored conditionally upon correct transcription in the earlier transcription phase. That is, the percent correct recall represents the percentage of sentences (for the first three methods) or words (for the Total Key Words method) correctly recalled that were also correctly transcribed during the study phase. For example, if a subject transcribed 43 sentences correctly by the Exact Words criterion, then the percent correct recall for Exact Words was calculated by dividing the number of exactly recalled sentences by 43. If, on the basis of the Key Words, 46 sentences were correctly transcribed, then the percent correct recall by Key Words was calculated by dividing the number of correctly recalled sentences by 46. The same method was used for scoring recall by Meaning. The rationale for this conditional scoring was that subjects could only correctly recall a sentence if they had correctly encoded it when they heard it during the initial study phase. If a stimulus sentence was transcribed incorrectly during study, we assumed it could not be "recalled" during the subsequent recall phase.

The results of each part of Experiment I will first be discussed separately, then the two parts will be compared. Experiment I-A was analyzed as a 2 X 2 factorial design with factors of Probe Cue (Visual or Auditory) and Talker (Single or Multiple). The data from List A and List B were combined within each condition for analysis, so that there were 96 sentences in each condition, half of which any given subject heard. A separate ANOVA was conducted on the three dependent measures of Exact Words, Key Words, and Meaning for transcription and probed recall. An alpha level of .05 was used for all statistical tests.

Experiment I-A: Transcription

Figure 1 shows the results of the transcription phase for each method of scoring in Experiment I-A. A slight increase in mean scores is observable as the scoring criterion changes from the most strict (Exact Words) to the least strict (Meaning). Within each of the three measures, it can be seen that the transcription performance in the single-talker condition was consistently better than in the multiple-talker condition.

The effect of talker variability was significant by all three scoring criteria: Exact Words, $F(1,76) = 6.296, p < .05$; Five Key Words, $F(1,76) = 9.08, p < .05$; and Meaning, $F(1,76) = 12.73, p < .05$. The effect for probe cue was not significant by any scoring criterion. (This was expected, since the study phase did not employ any cues. The factor of Probe Cue was included in the analysis of the transcription merely to establish that the conditions of Visual Cue and Auditory Cue were comparable during the transcription task.) The interaction between talker and cue was not significant by any method of scoring.

Experiment I-A: Probed Recall

The mean scores for the three dependent measures of the probed recall phase in Experiment I-A are shown in Figure 2. As with the transcription data, it can be seen that the mean scores increased slightly as the scoring criteria changed from the most strict to the least strict. On initial inspection, the distribution of recall scores in all conditions was found to be highly positively skewed. Therefore, a square root transformation of the raw scores was carried out to normalize the distributions prior to statistical analysis. Although Figure 2 shows the raw scores expressed as percent correct recall, the statistical analyses were performed on the square root transformation of these scores.
Figure 1. Percent correct sentence transcription from Experiment 1A, scored by Exact Words, Key Words, and Meaning.
Figure 2. Percent correct probed recall from Experiment I-A, scored by Exact Words, Key Words, and Meaning.
As can be seen in Figure 2, the mean recall scores in the visual cue conditions were greater for the single-talker lists than for the multiple-talker lists across all three scoring methods. However, in the auditory cue condition the difference between the single- and multiple-talker lists depended on the scoring criterion: when scored by Exact Words or Key Words, the mean scores were virtually identical for single and multiple talkers. When scored by Meaning there was an advantage for the single talker over the multiple talkers in both the visual and auditory cue conditions. Although these trends are apparent in the data from an examination of Figure 2, an analysis of variance failed to establish statistically reliable effects for any of the dependent variables. There were no statistically significant interactions between Talker and Cue. A t-test for independent samples showed a marginal difference between the means of the single-talker and multiple-talker conditions of the Visual Cue when scored by Exact Words, \( t(38) = 1.768, p = .085 \).

**Experiment I-B: Transcription**

Figures 3 and 4 show the results of the study phase of Experiment I-B, in which instructions about the recall part of the experiment were given to subjects at the beginning of the experiment. The transcription results replicate those of the study phase of Experiment I-A. Again, mean scores tend to increase in all conditions as the scoring criteria moves from most to least strict. (In terms of strictness, the fourth method added in this experiment, Total Key Words, falls between Key Words and Meaning.) The factor of Talker was highly significant by all four scoring criteria: Exact Words, \( F(1,76) = 14.062, p < .001 \); Five Key Words, \( F(1,76) = 10.069, p < .005 \); Total Key Words, \( F(1,76) = 16.391, p < .001 \) and Meaning, \( F(1,76) = 8.26, p = .005 \). Again, as expected, the factor of Probe Cue was not significant by any method of scoring (because cues were not used during transcription), nor was the interaction of talker and cue significant.

Insert Figures 3 and 4 about here

**Experiment I-B: Probed Recall**

Figures 5 and 6 present the means for the probed recall in each condition of Experiment I-B. Again, the means increased across the four methods of scoring in a manner similar to the transcription phase of this experiment and both phases of Experiment I-A. As with Experiment I-A, the means for the multiple-talker condition were lower than the single-talker condition when a visual probe cue was given for recall. When an auditory cue was given, however, the means of the multiple-talker conditions were higher than those of the single-talker conditions. This pattern is consistent across all four methods of scoring. For statistical analysis, the raw scores were transformed to their square roots, as in Experiment I-A. However, ANOVAs conducted on the transformed scores of the four dependent measures revealed no statistically significant main effects or interactions.

Insert Figures 5 and 6 about here

**Combined Analysis**

After these analyses were completed, the recall data from the two parts of Experiment I were combined and analyzed as a 2 X 2 X 2 design with factors of Talker (single or multiple), Cue (visual or auditory), and Instructions. This analysis was carried out for two reasons. First, this allowed us to compare the effect of incidental encoding of the stimuli in Experiment I-A to the deliberate encoding in I-B. This comparison could determine whether encoding of indexical information is automatic or strategic.
Figure 3. Percent correct sentence transcription from Experiment I-B, scored by Exact Words, Key Words, and Meaning.
Figure 4. Percent correct sentence transcription from Experiment 1-B, scored by Total Key Words.

- Auditory Cues
- Visual Cues

Multiple Talkers
Single Talker

Percent Correct
Figure 5. Percent correct probed recall from Experiment I-B, scored by Exact Words, Key Words, and Meaning.
Figure 6. Percent correct probe recall from Experiment 1-B, scored by Total Key Words.
(Geiselman, 1979; Geiselman & Bellezza, 1976, 1977; Geiselman & Crawley, 1983; Mullennix & Pisoni, 1990; Palmeri et al., 1993). Second, given the marginal results of the recall phase in both experiments, it was hoped that the greater power of the combined experiments might better show the effect of the probe cue or the talker condition. Only the first three scoring methods of Exact Words, Key Words, and Meaning were analyzed in the combined data, because Experiment I-A was not scored by Total Key Words. Recall was greater in Experiment I-B than Experiment I-A: in the combined recall data, a significant effect of Instructions was obtained across all three methods of scoring: Exact Words, $F(1,152) = 5.152, p < .05$; Key Words, $F(1,152) = 5.376, p < .05$; and Meaning, $F(1,152) = 16.744, p < .001$. When scored by Exact Words, the interaction of Cue and Talker was marginally significant, $F(1,152) = 3.195, p = .076$. When scored by Key Words, the factor of Cue was marginally significant, $F(1,152) = 3.214, p = .075$ and the interaction of Cue and Talker was also marginally significant, $F(1,152) = 2.844, p = .094$.

**Discussion**

**Transcription**

The transcription data from both parts of Experiment I provide evidence that the indexical properties of a talker's voice play an important role in the on-line perceptual processing of spoken sentences. Sentences from single-voice lists were consistently transcribed more accurately than sentences from multiple-voice lists. This finding is consistent with similar results concerning the perceptual processing of smaller units of spoken language such as phonemes and words (Jusczyk, Pisoni, & Mullennix, 1992; Mullennix et al., 1989; Mullennix & Pisoni, 1990; Nygaard et al., 1994; Sommers et al., 1992; Summerfield & Haggard, 1973; Verbrugge et al., 1976). The present findings suggest that processing sentences produced by multiple talkers requires more attentional resources or additional normalization processes compared to processing sentences produced by single talkers. These increased attentional resources appear to reduce the efficiency of initial perceptual processes for multiple-talker lists of stimuli compared to single-talker lists. This interpretation will be discussed in more detail in the general discussion.

**Probed Recall**

The probed recall data were the primary interest of the present set of experiments. Although no significant results were obtained in the recall phase of the Experiments I-A and I-B, several interesting trends were uncovered. In general, we observed a trend toward an interaction between Talker and Probe Cue. In both Experiments I-A and I-B, recall in the single-talker condition appears to be unaffected by the manipulation of the recall probe, whereas recall in the multiple-talker condition appears to be better with the auditory probe than with the visual probe. Between the two parts of the experiment, the magnitude of the increase in recall with the auditory cue for the multiple-talker condition is larger in Experiment I-B than in Experiment I-A. Furthermore, recall performance in all conditions is significantly better in Experiment I-B than in Experiment I-A. Thus, the explicit instructions to commit the sentences to memory in Experiment I-B did improve subjects' performance on the probed recall task, perhaps due to a greater depth of processing in Experiment I-B (Craik & Lockhart, 1972; Craik & Tulving, 1975).

Several possible explanations for these results are worth discussing. First, subjects' recall performance may be related to the perceptual effects of talker variability that are shown by the transcription data. As proposed by Martin et al. (1989), it may be that the early perceptual costs incurred in processing multiple voices are simply "passed on" to later stages of processing. For example, resources

---

3As discussed in the methods section, the mean intelligibility of the multiple-talker sentences was equivalent to that of the single-talker sentences when, in a previous study, listeners heard all stimulus sentences used in this study presented in single-talker lists.
in working memory may be of limited capacity, and the greater perceptual demands of multiple voices may reduce the amount of resources available in the later stages of working memory. As a result, the transfer of items from working memory to long-term memory may suffer in terms of efficiency, speed, or both. The representation of items in long-term memory might then be less robust for items in the multiple-talker condition than for items in the single-talker condition. We will refer to this explanation as the "perceptual encoding account."

Another possibility is that working memory is directly affected by the difference between the single-talker and multiple-talker stimuli. We will call this the "working memory account." Whatever the perceptual consequences of talker variability, it may be that attentional resources or rehearsal processes in working memory are allocated less efficiently when the items in working memory vary along a salient dimension. As mentioned above, this decreased efficiency or quantity of resources may result in less robust representations in memory for variable stimuli. This explanation receives support from the earlier study of Goldinger et al. (1991), who proposed that talker variability affects both early perceptual operations and working memory. The data from the visual cue conditions of both experiments are consistent with both the perceptual account and the working memory account. Both explanations predict that highly variable stimuli will be represented in long-term memory less robustly than stimuli of low variability. This is what we find in the visual cue conditions—lists of sentences spoken by multiple talkers are recalled less accurately than lists spoken by multiple talkers.

The data from the auditory cue conditions, however, are inconsistent with both of these explanations. In Experiment I-A, recall in the multiple-talker/auditory cue condition is equivalent to recall in the single-talker/auditory cue condition, whereas in Experiment I-B it is greater. This pattern of results suggests that the talker-specific information in the auditory cues is utilized by recall processes to identify the multiple-talker stimuli in long-term memory. Thus, the auditory cue condition suggests that multiple-talker stimuli are represented either equivalently to single-talker stimuli (in Experiment I-A) or more robustly (in Experiment I-B) than single-talker stimuli. Thus, neither the perceptual encoding account nor the working memory explanation as discussed here are consistent with the present data from both the visual and auditory cue conditions.

The results of the present study may be best explained if we retain the possibility that either the perceptual account or the working memory account (or both) may accurately explain the visual cue data, and we include an explanation at the stage of retrieval of items from long-term memory; such a retrieval account can accommodate the auditory cue data as well. A retrieval account rests on the proposal that the probability of retrieval of an item from long-term memory in a probed recall task depends on the match between the probe and the stored item. Specifically, the probability of retrieval is proportional to the number of matching cues between the probe and the stored item (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981). In our experiment, the variability of the multiple-talker lists of sentences may provide an additional contextual dimension that may be matched with probe cues during retrieval from long-term memory. Thus, in Experiment I-A, recall with the visual cues is better for the single-talker condition than the multiple-talker condition, perhaps because of the perceptual costs during initial processing or because of a deleterious effect of talker variability on working memory. However, when the auditory probe cues are provided, the talker-specific voice information serves as an additional retrieval cue or source of context that may be matched to items in memory. In the multiple-talker condition, this additional dimension may help reduce the search space in long-term memory to only those sentences that were spoken by the talker whose voice is represented in the probe cue. In contrast, the single-talker auditory probe cues would provide no additional information over that supplied by the single-talker visual probe cues, because the voice
information in the cue does not limit the search space in long-term memory; the voice information in the probe cue is consistent with all sentences in long-term memory.

Tulving and Thomson's (1973) encoding-specificity principle may apply here as well. In general, Tulving and Thomson found that retrieval of items from memory is best when the properties of items in memory—such as their modality and linguistic content—are matched by the properties of the retrieval cue. In the present experiment, the match between the auditory operations of encoding during study is better with the auditory retrieval cues than with the visual retrieval cues. Both of these retrieval accounts may explain why the recall performance does not change with the probe cue in the single-talker condition. In the multiple-talker condition, on the other hand, the voice information in the auditory probe cue appears to compensate for the perceptual costs or the decreased efficiency of operations in working memory during study.

EXPERIMENT II

The results of Experiment I suggest that the Talker and Probe Cue variables might produce effects on memory for spoken sentences that are important but not statistically detectable with the present methodology we employed. To pursue this possibility, we conducted another experiment in which Probe Cue was treated as a within-subject variable. (Recall that in Experiments I-A and I-B, all variables were between subjects.) Because of the complexity of the experiment, we could not also make the Talker variable within-subject. We also used shorter lists of sentences, which we expected would increase the recall scores over the generally low levels of recall performance in Experiment I.

Method

Subjects

Subjects were 80 undergraduate students (54 females and 26 males) enrolled at Indiana University who were paid $5.00 for their services. None reported any history of a speech or hearing impairment at the time of testing. All subjects were between the ages of 17 and 40 and all were native speakers of English.

Stimuli

The stimulus materials used in this experiment were taken from the same database of 2,000 sentences used in Experiment I. However, different sentences were used in Experiment II. Four sets of stimulus lists were compiled: Lists A and B spoken by a single talker; and lists A and B spoken by multiple talkers. The lists contained 28 sentences each. The single talker was the same as in Experiment I. The multiple-talker list contained 14 different talkers, each of whom spoke two sentences. As in Experiment I, the sentences were chosen so that the mean and variance of the intelligibility scores would be equivalent for both the single-talker and multiple-talker conditions. With the shorter sentence lists, it was possible to equate them precisely; the mean score for each of the four lists was exactly 9.50 and the standard deviation was exactly 0.51.

Procedure

The instructions given to subjects for this Experiment were similar to those of Experiment I-B, except they informed subjects that there would be two cycles to the experiment, a study phase followed by a test, then another study phase followed by another test. Subjects were not given the three suggestions about how to remember the sentences, as they were in Experiment I-B. Instead, they were simply told to "listen very closely to the sentences." This point was emphasized in the instructions. In all other respects, the instructions were identical to those used in Experiment I-B.
The experimental session was similar to that of Experiment I-B, except that subjects performed the study-test cycle twice, once with an auditory probe cue and once with a visual probe cue. They received a different list of sentences for each cycle. Subjects were randomly assigned to conditions, and the order of the cues and of the sentence lists was counterbalanced in all conditions. The talker variable was manipulated between subjects. There were 26 experimental sessions, with one to five subjects in each session.

Results

Transcription

As expected, there was no effect of the probe cue in the transcription phase. Figure 7 shows the results of the transcription phase for the four methods of scoring used for Experiment I-B. Since we had firmly established the equivalence of the probe cue condition in the transcription phase, Figure 7 shows the transcription results collapsed across probe cue conditions.

The transcription data are quite similar to those found in Experiments I-A and I-B. Again, the scores increased as the scoring criteria changed from most strict to least strict. Most importantly, we found the same distinct advantage in transcription performance for the single-talker condition across all scoring methods. This talker effect was significant by all scoring criteria: Exact Words, $F(1,152) = 3.392, p < .05$; Five Key Words, $F(1,152) = 6.810, p < .05$; Meaning, $F(1,152) = 13.643, p < .001$; and Total Key Words, $F(1,152) = 17.856, p < .001$.

Probed Recall

The recall data are presented in Figure 8. A clear trend can be seen for the percent correct recall to be higher in the multiple-talker condition than in the single-talker condition across all scoring criteria.

As in Experiment I-A, the data were found to be skewed positively upon inspection. Therefore, a square-root transformation on the raw scores was performed to normalize the distributions. Although Figure 8 shows the raw data, the analysis was conducted on the transformed data. A significant main effect for Talker was found by three scoring criteria: Exact Words, $F(1,152) = 4.181, p < .05$; Five Key Words, $F(1,152) = 5.886, p < .05$; and Total Key Words, $F(1,152) = 4.625, p < .05$. The main effect of Probe Cue was not significant by any scoring criterion, nor was the interaction of Talker and Cue.

Discussion

Transcription

The transcription data from Experiment II replicate the results of Experiments I-A and I-B. Transcription accuracy was significantly lower for multiple-talker lists of spoken sentences than for single-talker lists by all four methods of scoring (Fig. 7). These data support earlier findings that listeners incur processing costs as they accommodate changes in the voice of spoken words and phonemes in an
Figure 7. Percent correct sentence transcription from Experiment II, scored by Exact Words, Key Words, Meaning, and Total Key Words, collapsed over Visual and Auditory Cue conditions.
Figure 8. Percent correct probe recall from Experiment II, scored by Exact Words, Key Words, Meaning, and Total Key Words.

- Multiple
- Single

Percent Correct

0 5 10 15 20 25 30 35 40 45 50 55 60
identification or transcription test (Jusczyk et al., 1992; Mullennix et al., 1989; Mullennix & Pisoni, 1990; Nygaard et al., 1994; Sommers et al., 1992; Summerfield & Haggard, 1973; Verbrugge et al., 1976). Possible explanations of these results will be discussed in detail in the general discussion.

**Probed Recall**

Given the recall results of Experiments I-A and I-B, the recall data for Experiment II were quite unexpected. Overall, probed recall was significantly higher for multiple-talker lists across both visual and auditory probe cues (Fig. 8). However, there was no hint of an effect of probe cue, nor of an interaction between type of Probe Cue and Talker. It is difficult to reconcile these results with those of Experiments I-A and I-B. They may be due to the several procedural changes made between Experiment I and Experiment II. One of these changes was that the factor of Probe Cue was made a within-subjects manipulation in Experiment II, whereas it had been a between-subjects manipulation in Experiments I-A and I-B. Our motivation for this change was the hope that making it a within-subject factor would reduce some of the variation between scores due to variation between subjects. It seems unlikely that this change would have caused the effect of cue to be less significant here than it was in Experiment I.

Another change in Experiment II was that subjects heard shorter lists of sentences in each condition, 28 in Experiment II compared to 48 in Experiments I-A and I-B. This change was made so that subjects could be assigned to two conditions of the experiment (visual and auditory cues) without significantly extending the time necessary for an experimental session. We expected this change to increase recall performance in general, but we were uncertain of what effect it might have on the factors of Talker and Probe Cue. As can be seen in Figures 2, 5, and 6, recall in the previous experiments was most difficult in the visual cue/multiple-talker condition. Comparing these two figures to Figure 8, it can be seen that the effect of the changes in Experiment II was most pronounced in this condition. Perhaps the shorter lists produced a relatively greater advantage in the visual cue/multiple-talker condition than in the other conditions.

Another possibility is that subjects learned something about the task during the first study-test cycle, and then changed their encoding strategies for remembering the stimuli for the second study-test cycle. However, any such effect should be minimal on the overall results, since the order of probe cue was counterbalanced in all conditions. To verify that the order of conditions was unimportant, we conducted a post-hoc analysis of variance that included a factor of Order of Probe Cue. This factor was significant only when recall was scored by Meaning ($F(1,152) = 5.427$, $p = .021$). There were no interactions with other factors. (As discussed above, scoring by Meaning is the most subjective criteria, and we believe it is less revealing of subjects' performance than the other three criteria.) We conclude that the effect of including two cycles within an experimental session is minimal. In general, we are unsure why the Probe Cue should have had no effect in Experiment II. However, since its effects in Experiments I-A and I-B were not significant, we will confine the rest of the discussion to the effect of Talker.

The factor of Talker in this experiment also conflicts with its apparent effect in Experiments I-A and I-B. Here, we found a main effect of Talker, while in Experiments I-A and I-B we did not. However, because the results from Experiment II were statistically significant, whereas those from Experiments I-A and I-B were not, we will not attempt a reconciliation of the two. We will simply offer several possible explanations for the effect of talker found in Experiment II. Two accounts may be offered that rely on differences in attentional processes during the perceptual processing of the stimuli. First, the change of voices from trial to trial might simply draw attention to the stimuli to a greater degree than when the voices do not change. This greater attention might cause the multiple-talker sentences to be encoded more robustly into long-term memory. A second attention-based possibility is that, because the stimuli in the multiple-
talker condition are more difficult to process perceptually, subjects must rely upon the linguistic content and structure of the sentences to constrain possible interpretations more than they do in the single-talker condition. Attending to the semantic and syntactic information in the sentences may, in turn, encourage a deeper level of processing (Craik & Lockhart, 1972; Craik & Tulving, 1975) than is required in the single-talker condition. This also may result in a more robust representation in memory of the multiple-talker sentences.

The effect of Talker may also be due to processes that operate during recall itself. Again, we offer two proposals. First, due to the encoding in long-term memory of talker-specific information, the multiple-talker sentences may be more distinctive in memory than the single-talker sentences. The additional dimension of voice which differentiates the stimuli in the multiple-talker condition may provide an additional retrieval cue that can be used during retrieval. Such processes may be related to the well-known advantage in recall for phonologically dissimilar items (e.g., Baddeley, 1966a, 1966b; Conrad, 1964; Conrad & Hull, 1964; Wickelgren, 1965). Second, it is possible that the recall of a particular sentence may inhibit other sentences that share dimensions with the retrieved sentences. Thus, in the single-talker condition, when a sentence is activated, all the sentences spoken by that same talker may be inhibited. However, in the multiple-talker condition, the activation of a particular sentence would only inhibit one other sentence in memory, since only two sentences in this condition were spoken by a given speaker.

All four of these accounts—two attention-based and two retrieval-based—are consistent with the study of Goldinger et al. (1991), who found that when the rate of presentation of lists of multiple-talker words was reduced, recall for multiple-talker words actually surpassed that of single-talker words in the primacy portion of the serial recall list. Because the duration of sentence-length stimuli in our experiment is much greater than the word-length stimuli used by Goldinger et al., and because sentences were presented very slowly (20 seconds apart) in all of the present experiments, we might expect such an advantage for the multiple-talker stimuli in our experiment as well.

Concerning the different methods of scoring the transcription data, a comparison of subjects' performance across the four criteria shows that the different criteria reflect very similar aspects of the subjects' performance. In each case, the transcription scores are higher in the single-talker condition than in the multiple-talker condition. The scores generally increase as the criteria change from the most strict (Exact Words) to the least strict (Meaning). In Experiment I-B, the addition of the Total Key Words criterion also appears to reflect the same aspect of subjects' performance. From this data, there is no apparent reason for using one method over another.

The several different ways of scoring show similar patterns in the recall data, with one small exception. The three form-based methods of scoring (Exact Words, Key Words, and Total Words) show the same patterns, while scoring by Meaning gives a slightly different result in Experiment I-A than in I-B or II. Scored by the three form-based criteria, recall performance in the multiple-talker condition was consistently higher with the auditory cues than with the visual cues. However, scored by Meaning, there is only a very slight increase in scores with auditory cues. This slight difference in the data when scored by Meaning does not appear to require a different interpretation of the data. Overall, the transcription and recall data suggest that the three form-based methods of scoring are essentially equivalent, and any single one could be used. Because the Meaning criterion is much more subjective on the part of the scorer, and does not appear to illuminate any additional aspect of the subjects' performance, we conclude that it may be dispensed with in the present discussion and in further studies.
GENERAL DISCUSSION

Taken together, the data from all three experiments provide preliminary evidence that talker variability affects the perception of spoken sentences in a manner that is similar to the way it affects the perception of isolated spoken syllables and words. With respect to the neural representation of spoken sentences in memory, the present results are, unfortunately, inconclusive at the present time. Both abstractionist and non-analytic accounts of perception are consistent with the recall data from these experiments. Nevertheless, the present findings provide several new insights into the encoding and retrieval processes employed in the perception of spoken sentences. In the sections below, we will first discuss the significance of the transcription data and their relevance to current theoretical issues in speech perception. Then we will consider several interpretations of the recall data. Finally, we will offer some suggestions for future research.

Transcription Data

The transcription task used during the study phase in all three experiments was identical and the results were generally quite consistent. Transcription accuracy was consistently lower in the multiple-talker conditions than in the single-talker conditions. Thus, there appears to be a perceptual cost to a trial-to-trial change in the voice of spoken sentences. Although such perceptual costs of talker variability have been found for isolated words and phonemes (e.g., Fourcin, 1968; Mullemix et al., 1989; Rand, 1971; Verbrugge et al., 1976; Weenink, 1986), one might expect that the much greater constraints on acoustic-phonetic processing of sentences, imposed by their syntactic and semantic structure (Miller, Heise, & Lichten, 1951), would considerably reduce if not eliminate the effects of surface variation on perception. Indeed, words in sentences are widely known to be more intelligible than words in isolation (e.g., Grosjean, 1980; Grosjean & Gee, 1987; Marslen-Wilson & Welsh, 1978; Miller et al., 1951; Miller & Isard, 1963; Morton & Long, 1976; Rubenstein & Pollack, 1963; Spoehr, 1980). Consequently, it was somewhat surprising to find that the effects of stimulus variability on spoken sentences were similar to those observed for isolated phonemes and words. Our surprise is similar to that of Nusbaum and Morin (1992) who thought listeners' knowledge of lexical structure might attenuate the effects of talker variability on words as compared to isolated vowels. However, they found variability to affect recognition of words more than recognition of vowels. The present results provide evidence that, even given the highly constraining semantic and syntactic structures of a sentence, indexical attributes of the speech waveform significantly affect perceptual processing in a manner that is similar to the way they affect perceptual processing of spoken words and phonemes. To our knowledge, these results provide the only replication at the sentence level of Peters' (1955) little-known study on the perception of sentence-length messages spoken by multiple talkers.

Since the effects of talker variability on the processing of spoken sentences appear to be similar to its effects on isolated words and phonemes, theories that have been developed to explain the perceptual consequences of talker variability on isolated words and phonemes are relevant to the present study. As discussed earlier, the general finding on isolated words and phonemes is that single-talker lists are perceived more accurately and/or faster than multiple-talker lists (e.g., Assmann, Nearey, & Hogan, 1982; Creelman, 1957; Mullemix et al., 1989; Strange & Gottfried, 1980; Strange, Verbrugge, Shankweiler, & Edman, 1976; Summerfield, 1975). Many theorists assume that the perceptual mechanisms responsible for these differences in perceptual processing are related to differences in the mapping of the acoustic information onto phonemic categories across multiple talkers. Differences in this mapping have long been known to speech researchers as the problem of perceptual constancy (Goldinger et al., 1995; Klett, 1986; Liberman, 1957; Liberman et al., 1967; Shankweiler et al., 1976). Although the acoustic properties of speech vary considerably from talker to talker, listeners perceive the linguistic value of the speech signal
with no apparent difficulty. The question of how this is accomplished has continued to occupy researchers for many years. In a recent review article, Nearey (1989) distinguishes two main classes of theories that have attempted to explain the perception of vowels. Nearey adopts the terminology first proposed by Ainsworth (1975) and refers to these as extrinsic and intrinsic theories of vowel perception. We will discuss the present results from the sentence transcription task in terms of each of these theories.

Vowel identification has been assumed by many researchers to be primarily dependent upon the first two energy peaks of the steady-state spectrum of a vowel (Assmann et al., 1982; Chiba & Kajiyama, 1941; Joos, 1948; Ladefoged, 1967; Miller, 1989; Nearey & Assmann, 1986; Peterson, 1961). The frequencies of these formants can be used to distinguish many vowels; however, there is also substantial overlap in F1 x F2 space for several vowels (Ladefoged & Broadbent, 1957; Peterson & Barney, 1952). An extrinsic solution to this perceptual constancy problem involves a rescaling of a particular speaker's vowel space so that such overlap is reduced or eliminated (e.g., Gerstman, 1968; Lieberman, 1973). In this "self-normalization" model, a listener uses an extended sample of a given speaker's speech to map the boundaries of his or her vowel space. By this process, the listener learns to evaluate the specific formant frequencies of a given talker in terms of standard vowel categories. Although the concept of extrinsic normalization has most often been applied to vowel perception, J.L. Miller (1987) has shown that listeners also use individuals' rates of speech to calibrate the voice-onset time (VOT) of stop consonants (see also Miller & Liberman, 1979).

The data from the transcription task used in all three of our experiments are consistent with an extrinsic account of speaker normalization. By this account, this recalibration or rescaling has to be carried out only once for the entire single-talker condition. However, in the multiple-talker condition, it has to be done on every trial, each time the speaker changes. If this recalibration is not done immediately on the first segment of speech, but instead requires several samples from each talker (the necessary sample size has not been investigated by researchers), listeners' perceptions may be relatively inaccurate until this recalibration is complete. Thus, the reduced efficiency of the initial perceptual analysis may account for the reduced transcription accuracy in the multiple-talker condition.

We can contrast this explanation with an "intrinsic" account of perceptual normalization, in which the information required for recalibration or rescaling is assumed to be contained within each segment or syllable of speech (Nearey, 1989; Nusbaum & Morin, 1992; Strange, 1989a; Syrdal & Gopal, 1986). In some versions of this view of normalization, recalibration is accomplished by comparing the F1 and F2 frequencies of vowels, which vary considerably across vowels, to the F0 and/or F3 frequencies, which are relatively stable across vowels. These values are often assumed to be extracted from stable portions of vowel segments. In other versions of this view, dynamic information, such as duration and formant transitions into and out of consonants, is assumed to provide invariant cues to vowel identity (Rakerd & Verbrugge, 1987; Strange, 1989a). In either case, the problem of talker variability is considered to be illusory; the acoustic signal corresponding to individual syllables is believed to carry sufficient information for a listener to correctly categorize vowels across talkers without recourse to memory for prior instances.

At first glance, it would seem that our results and the earlier findings showing a perceptual disadvantage for trial-to-trial changes in the voice of stimuli (Fourcin, 1968; Mullemnix et al., 1989; Mullemnix & Pisoni, 1990; Rand, 1971; Summerfield & Haggard, 1973; Verbrugge et al., 1976; Weenink, 1986) are inconsistent with intrinsic accounts of normalization. An intrinsic account would predict no effect of changing the voice of stimuli across trials because an individual segment or syllable contains all of the necessary information for linguistic categorization (Nusbaum & Morin, 1992). However, it could be the case that the perceptual system must learn something about a given speaker's voice in order to accomplish this intrinsic identification most efficiently. Although Strange (1989a) does not explicitly state this
possibility, she does leave room for it in her reformulation of the problem of perceptual constancy in speech perception:

If, as the perceptual results suggest, there is sufficient information within single syllables to allow the listener to identify intended vowels, even when those vowels are coarticulated by different speakers in different consonantal contexts, then the need to postulate psychological processes by which the perceiver enriches otherwise ambiguous sensory input is eliminated. Instead, the goal becomes the development of perceptual models that describe how perceivers detect and utilize the available sources of information for vowel identity during the perception of ongoing speech. [italics added] (p. 2085)

This call for researchers to focus more on psychological processes associated with perceptual constancy than on traditional attempts to characterize the speech waveform is seconded by Neary (1989). He views extrinsic and intrinsic theories of vowel normalization as hypotheses about what it is in the speech waveform that listeners learn about when they gain experience with particular talkers. Such hypotheses, however, are not explanations of how listeners modify their perceptual operations for particular talkers. That is, theories of vowel normalization focus primarily on information in the acoustic waveform and say little about the perceptual or neural mechanisms that are actually employed to encode this information (Nusbaum and Morin, 1992). Neary (1989) refers to extrinsic and intrinsic accounts of normalization as "data analytic" theories and urges that they be wedded to psychological data on perception before they are considered to be relevant to theories of human speech perception.

Nusbaum and Morin (1992) conducted a series of experiments that were designed to evaluate the psychological claims of data-analytic accounts of perceptual normalization. First, they replicated several previous results (e.g., Creelman, 1957; Fourcin, 1968; Mulleinnix et al., 1989; Mullennix & Pisoni, 1990; Rand, 1971; Weenink, 1986). They found that isolated vowels, CV syllables, and words were all identified more slowly when they were presented in multiple-talker formats than when they were presented in single-talker formats. They also found that a digit-preload task increased the identification time of syllables and isolated vowels more in the multiple-talker condition than in the single-talker condition. Nusbaum and Morin interpreted this finding as a demonstration that adjusting to different voices from trial to trial is an attention-demanding perceptual operation (see also Luce, Feustel, & Pisoni, 1983). This attention-demanding nature of variability had been demonstrated by Martin et al. (1989) in a serial recall task; Nusbaum and Morin showed that the attentional deficits associated with talker variability are not unique to rehearsal processes in memory tasks but arise during perceptual identification as well. They proposed that these effects arise from an inconsistent mapping of acoustic information to linguistic responses when talkers' voices change from trial to trial, as has been found in research on attention in other cognitive tasks (e.g., Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977).

Finally, Nusbaum and Morin suggested that perceptual operations might be different between single-talker and multiple-talker conditions. To test this hypothesis, they removed either F0, F3, or both from isolated vowel stimuli, assuming that "structural estimation" (their term for intrinsic normalization) processes primarily rely upon these formats, as proposed by Syrdal and Gopal (1986). If structural estimation were used to different degrees in the multiple-talker and single-talker conditions, the absence of this information should affect the two conditions differently. Nusbaum and Morin found that the degree of degradation of the stimuli (missing either F0, F3, or both) affected accuracy of identification in the multiple-talker condition only. There was no effect observed in the single-talker condition. Nusbaum and Morin concluded that structural estimation is used primarily in the multiple-talker condition, but that "contextual tuning" (their term for extrinsic normalization) is used when the voice remains constant. We find this an interesting proposal concerning the question of how perceptual operations contend with talker
variability. It is possible that the differences in transcription accuracy between the present single-talker and multiple-talker conditions are due to differences in perceptual operations used in the two conditions.

Nusbaum and Morin's proposal that talker variability increases demands on attention during processing receives further support from a recent study of working memory by Saldaña (1995). Saldaña found that subjects who were presented with digits spoken by multiple talkers were able to keep fewer digits in working memory than subjects who were presented with digits spoken by a single talker. Thus, the greater variability of the multiple-talker stimuli appeared to require more resources in working memory than the single-talker stimuli.

Although the data from the present transcription tasks do not confirm or disconfirm Nusbaum & Morin's proposal of different perceptual mechanisms operating in the single-talker and multiple-talker conditions, they are consistent with this proposal. Furthermore, their proposal receives support from several prominent theories of perception and memory. Kolers (1973, 1976; Kolers & Roediger, 1984) argues that perception is subserved by operations that are specific to a given stimulus and that can be learned with practice. In one study, Kolers showed that the efficiency with which subjects processed inverted text increased with practice (Kolers & Ostry, 1974). A very similar idea, proposed by Jacoby & Dallas (1981), is that experience with stimuli enhances "perceptual fluency," i.e., the ease or efficiency with which stimuli are perceived or categorized. In the present case, a listener's experience with the way the acoustic waveform of a specific talker specifies a particular talker's vocal tract may enhance the fluency with which he or she can recognize speech from that talker.

Proposals of this sort have received support in the speech domain by Nygaard et al. (1994). Nygaard et al. showed that listeners are better at perceiving isolated words when they can identify the speakers by their voices than when they cannot identify the speakers. When listeners know a talker's voice well enough to identify it, they appear to know something about the talker-specific characteristics of the speech waveform that enhances the fluency with which they perceive the linguistic content of that voice—even with novel utterances from the talker.

Closely related to perceptual fluency is the idea of "transfer-appropriate processing" (Morris, Bransford, & Franks, 1977; Roediger & Blaxton, 1987). This refers to the performing of similar mental processes during study and test phases of an experiment. For example, Roediger and Blaxton (1987) had subjects study a list of words that was in one of several presentation forms: handwritten, typed-focused, or typed-blurred. Subjects were then presented with fragments of the studied words (e.g., _YS_E_Y for _mystery_). The fragment was either in the same or different form as the studied item. Performance was better when the form of the word-fragment matched that of the originally studied word. Although the present transcription task is different from a word-fragment completion task, the perceptual operations that underlie transfer-appropriate processing in word-fragment completion may also underlie the superior transcription accuracy in the single-talker condition of the present experiment. In the single-talker condition, subjects may identify linguistic segments of the speech signal better because they acquire experience with the particular form of those segments (i.e., the talker-specific waveform).

In sum, the data from the transcription tasks of the present experiments are consistent with several general theories of perception and memory such as learning of perceptual operations (Kolers, 1973, 1976; Kolers & Roediger, 1984), perceptual fluency (Jacoby and Dallas, 1981), and transfer-appropriate processing (Morris, Bransford, & Franks, 1977; Roediger & Blaxton, 1987). They are also consistent with Mullenix and Pisoni's (1989) proposal that talker variability imposes a cost on perceptual operations arising from increased demands on attentional processes or resources.
Probed Recall Data

Although the probed recall data are somewhat equivocal concerning the representation of indexical properties of speech in memory for spoken sentences, they do demonstrate that the effects of talker-specific information in spoken language are evident throughout the various stages of recall tasks—from perception, analysis, and storage, through retrieval and recall itself. Since no significant effects were obtained in the recall tasks of Experiments I-A or I-B, we will consider here only the data from Experiment II, where reliable differences were observed.

The primary finding from the recall data in Experiment II was that the sentences in the multiple-talker condition were recalled better than those in the single-talker condition (regardless of Probe Cue). This pattern contrasts sharply with the findings from the transcription task in which sentences in the multiple-talker condition were perceived less accurately than those in the single-talker condition. We will consider four possible interpretations of these results. Two of these explanations could be framed within either traditional symbolic accounts of perception and memory or within more recent non-analytical accounts; they make no explicit claim about whether talker-specific information in the speech signal is incorporated into the memory representations of the sentences. The second two interpretations we will consider will be consistent only with non-analytic accounts of perception and memory; they specifically propose that talker-specific information is encoded directly into long-term memory during perception.

We will first consider the two explanations that could be given from either traditional analytic theories or non-analytic theories of cognition. These two accounts focus on possible differences in perceptual operations between the single-talker and multiple-talker conditions that arise during the initial perception of the sentences rather than during recall processes. According to these views, recall operations operate equivalently in the two conditions. The difference in recall performance, then, simply reflects differences in encoding operations during perception, working memory, and/or transfer of the stimuli into long-term memory.

The first account of this type we offer is that the higher recall scores in the multiple-talker condition are due to a difference in attentional processes during perceptual analysis. It may be that the greater variability in the multiple-talker condition automatically recruits attentional resources to a greater degree than the less variable single-talker sentence set. Normalization processes that discard talker-specific information may indeed operate on the stimuli, but greater attention is focused on the multiple-talker sentences before they are converted into abstract segmental representations. This explanation receives some support from the findings of Goldinger et al. (1991), in which multiple-talker lists of spoken words were recalled better than single-talker lists when the inter-stimulus intervals were sufficiently long (i.e., greater than 2000 ms). Goldinger et al. employed a serial recall task in which subjects recalled lists of 10 words in the exact order in which they had been presented. Goldinger et al. proposed that the advantage in primacy recall for multiple-talker lists may have been due to more thorough elaboration or other attentional processes (Glanzer & Cunitz, 1966; Murdock, 1962) that, at long ISIs, could operate on the distinctive information in the multiple-talker sentences. At short ISIs, however, the additional information in the multiple-talker sentences may have distracted or overloaded subjects' attentional resources. In the present study, the ISI was 20 seconds, which is much longer than the longest ISI, 4 seconds, used in Goldinger et al.'s study with word lists. Therefore, subjects in the present study presumably had sufficient time for rehearsal and elaboration of the sentences. (Of course, our stimuli were also much longer than the isolated words used by Goldinger et al.; however, the ratio of the ISI to the stimulus length was also much greater.
in the present experiment. 4) The advantage for the multiple-talker lists in the present experiment therefore may be due to increased rehearsal or other attentional or coding processes during the transcription task. Again, such an explanation makes no specific proposal about whether talker-specific information is included as part of the long-term memory representation of the sentences. Consequently, this explanation could be accommodated within either traditional abstractionist theories of cognition—by proposing that talker-specific information is filtered out before the sentences are transferred to long-term memory—or within non-analytical theories of cognition—by proposing that talker-specific information is retained in the long-term memory representation of the sentences.

A second explanation that is consistent with either analytical or non-analytical accounts of perception is also possible. Rather than proposing a difference in attentional processes in the two conditions, this explanation proposes a difference in levels of processing between the single-talker and multiple-talker conditions. As shown by the transcription data, the increased variability in the multiple-talker condition reduces the accuracy of perceptual encoding relative to the single-talker condition. This relative difficulty of the multiple-talker condition may cause subjects to think more about the semantic content and syntactic structure of the sentences in order to resolve perceptual ambiguities encountered during processing. Because subjects think more about the meaning and structure of the sentences, they may in turn process the sentences more deeply (Craik & Lockhart, 1972; Craik & Tulving, 1975) and consequently encode them into long-term memory more effectively in terms of semantic cues and elaboration. As with the previous attentional account, this levels-of-processing account makes no commitment about the representation of perceptual details in memory; it is consistent with either an abstractionist or a non-analytic model of memory. For the moment, we simply note that it could be consistent with traditional analytic accounts.

Other possible explanations are inconsistent with analytical or symbolic accounts and are instead consistent with non-analytic accounts of perception that are becoming more influential throughout cognitive psychology (Jacoby & Brooks, 1984; Kirsner & Dunn, 1985; Tulving & Thomson, 1973; Whittlesea, 1987, 1993; Whittlesea & Brooks, 1988). By this view, talker-specific information is explicitly assumed to be encoded into the long-term memory representation of the sentences. Furthermore, the difference in recall between the two conditions is due to differences in the representation of the sentences in memory and in the operations on these representations that retrieve the sentences from memory.

The first non-analytic account we will offer is that the greater variability of the multiple-talker sentences in memory—due to the encoding of fine perceptual details of the stimuli into long-term memory—provides additional contextual cues for use in retrieval. That is, the voice information of the multiple-talker stimuli in memory provides an additional dimension of representation that may be used to differentiate these sentences in memory in a way that is not possible in the single-talker condition. Retrieval mechanisms may take advantage of this distinctiveness among the multiple-talker sentences. Note that by this account the presence or absence of voice information in the probe cues was irrelevant to subjects' performance. In both the visual and auditory cue conditions, the probe cue may have supplied only the linguistic cue for retrieval. Using this linguistic information, however, was more effective in the multiple-talker condition because of the greater distinctiveness of the sentences in memory in this condition. This proposal receives some support from studies showing that items that are similar tend to be recalled less accurately than items that are relatively more differentiated. Of particular relevance to the present study is the well-established finding that items that are phonologically similar in short-term memory are recalled

4The relatively large ratio of ISI to stimulus length in the present experiments may account for the lack of effect of the Probe Cue; it may be that the long rehearsal time was so effective for the elaboration of multiple-talker sentences that the effect of the cue was overpowered.
less accurately than items that are phonologically dissimilar (Baddeley, 1966a, 1966b; Coltheart, 1993; Conrad, 1964; Conrad & Hull, 1964; Nairne, 1990; Wickelgren, 1965). It may be that perceptual similarity displays similar effects in long-term memory, and that voice information is one dimension on which similarity of items is computed.

A second explanation for the present results that is consistent with a non-analytic account of perception is based on an interactive-activation model of retrieval of items from memory. It is possible that, as an item is identified and retrieved from memory, other items matching the retrieved item on relevant dimensions may be inhibited. This idea has been extensively employed in neural network models of memory in which connections between nodes at a given level of representation are inhibitory (e.g., McClelland & Elman, 1986; see also Neely, Schmidt, & Roediger, 1983; Roediger & Neely, 1982). If voice information is a dimension of the representation of sentences in memory, the activation of a given sentence in memory may inhibit other sentences spoken by the same speaker. Applying this idea to our results, the activation of a sentence in the single-talker condition may inhibit all other sentences in that condition, whereas the activation of a sentence in the multiple-talker condition would only inhibit one other sentence. (In the multiple-talker condition of Experiment II, each talker spoke two sentences.) Depending on the decay rate of inhibition, fewer sentences in the multiple-talker condition may be inhibited on average during recall, leading in turn to an advantage in recall in the multiple-talker condition.

Four different explanations have so far been discussed to account for the recall data, each proposing a specific difference in cognitive operations between the single-talker and multiple-talker conditions: (a) a difference in attention, affecting perceptual and rehearsal processes; (b) a difference in levels of processing, also affecting perceptual operations; (c) a difference in distinctiveness of stimuli in long-term memory, affecting retrieval operations; and (d) a difference in inhibition, also affecting retrieval operations. As noted above, the first two of these explanations are consistent with both symbolic and non-analytical accounts of cognition, whereas the second two are consistent only with non-analytic accounts. It is worth noting here that these possible explanations are not mutually exclusive—a combination of them may in fact be responsible for the present results. The combinations that are possible depend on the resolution of the debate between analytic and non-analytic cognition. Specifically, if non-analytic cognition is assumed, only the first two accounts discussed above may be candidate explanations of the present results. However, if non-analytic cognition is assumed, then any combination of the four accounts is possible. For example, they may all contribute to the differences in recall. Consider the effects of talker variability in each of the above accounts: it may increase attention to the highly variable stimuli during perception; it may foster a deeper level of processing on these stimuli; it may result in greater distinctiveness of these stimuli in memory which may aid in retrieval; and it may inhibit other stimuli less during retrieval than the less-variable single-talker sentences. We recognize that such a multi-faceted explanation is rather unattractive in terms of theoretical parsimony at the present time. However, tentatively adopting the possibility that each of these factors may play a role in determining the present results does recognize the complexity of the operations that are generally assumed to underlie the perception and recall of complex stimuli such as spoken sentences. Further research is currently underway in our laboratory to pursue these important questions.

Although our results do not support the original hypothesis that the presence or absence of voice information in the probe cue would affect recall performance, it is important to note that the results are consistent with the proposal that talker-specific information is represented in the long-term memory representation of spoken sentences. Furthermore, given the bulk of evidence presented in the introduction suggesting that phonemes, syllables, and words are processed in a non-analytic manner, we conclude that
the same is true of spoken sentences. Of course, we regard this conclusion as preliminary, and await further results to help resolve these issues.

We believe there are several ways the present experimental paradigm could be modified to provide more information to answer these questions. First, we plan to use longer probe cues in the recall task. It may be that we found no reliable results of the visual cue versus auditory cue manipulation because the cues were very short, in some cases consisting of only a single word. Perhaps such short samples of speech do not provide sufficient talker-specific information to be used by retrieval mechanisms. Using three or four key words as probe cues may allow listeners to more effectively extract talker-specific information from the cues. If this information were then used as an additional retrieval cue in the auditory cue condition, and recall were greater for the auditory cues than for the visual cues, we would have clear evidence that the representation of sentences in memory does contain talker-specific information. This is the next experiment we will carry out.

A variety of other modifications of the present experimental paradigm might be helpful as well. It may be informative to explicitly vary the depth of processing (Craik & Lockhart, 1972; Craik & Tulving, 1975) during the study phase to determine whether the effect of talker-specific information on recall varies with depth of processing during study. Some tasks that would promote shallow processing of the sentences would include having subjects monitor the sentences for a particular word, count the number of nouns in a sentence, or answer other questions based on "superficial" aspects of the stimuli. Deeper levels of processing might be fostered by requiring subjects to answer specific questions about the meaning of the sentences. One obvious manipulation of levels of processing would be to vary subjects' explicit attention to the voice of the stimuli. Subjects could be required to attend to the voice by using a continuous recognition paradigm (such as that used by Palmeri et al., 1993) in which they must decide for each voice whether they had heard it previously. To manipulate attention away from the voice of the sentences, they could instead judge whether a given sentence itself, regardless of the speaker, had been presented previously. It might also be helpful to use other memory tests as converging operations on these issues. Instead of cued recall, recognition memory could be assessed with tasks requiring subjects to judge whether or not a sentence presented at test was previously presented. If talker-specific information is retained in memory, we would expect that same-voice repetitions of sentences would be recognized more accurately than different-voice repetitions (as has been found for isolated words; e.g., Palmeri et al., 1993). A continuous-recognition memory experiment is currently underway to test this prediction.

In summary, the present research provides preliminary evidence that stimulus variability is an important factor in both encoding and recall of spoken sentences. Although several interpretations of the present findings are possible, our results are consistent with non-analytic theories of perception and memory. It is likely that the memory representation of spoken sentences, like that of isolated words and phonemes, consists of a detailed perceptual trace of the original acoustic signal. Indeed, spoken sentences may provide an important stimulus domain in which to test non-analytic theories of cognition and learn more about the role of indexical and paralinguistic attributes in speech perception and spoken word processing.
References


