Talker-Specific Effects in Recognition Memory for Sentences

Kipp McMichael

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

1 This work was supported by a grant from the NIH-NIDCD Research Grant DC00111. For his detailed, insightful, and exhaustive editing of this paper, I thank David Pisoni. Special thanks also goes to Luis Hernandez for his technical assistance during the completion of this research.
Talker Specific Effects in Recognition Memory for Sentences

Abstract. Speech perception is usually considered to be the process by which listeners change spoken sounds into a meaningful string of words and ideas. Yet information about the talker’s age, gender, and socioeconomic status is also carried in parallel with the symbolic content of the linguistic message. The traditional view of speech perception posits that generic linguistic units such as phonemes and words are recovered from the speech signal while “extra-linguistic” talker-specific attributes such as gender, dialect, and emotional state are filtered from the signal during perceptual processing and before encoding into memory. The last decade has seen a dramatic increase in speech research specifically designed to measure the effects of stimulus variability. Most of this earlier work has used spoken word recognition to study speech processing, encoding and recall. The present series of four recognition memory experiments utilized sentences as stimuli to examine the process of sentence encoding and later recognition. Subjects listened to a study phase of 40 sentences spoken by 5 male and 5 female talkers and then completed a recognition test phase of 80 sentence which were a mixture of study phase and unstudied talkers and sentences. Results of all four experiments revealed significant effects of voice on sentence judgment accuracy and discrimination scores. These results call into question current theories and models of speech processing which posit preliminary normalization or other variability reduction. While no current theory models speech recognition outside a framework incorporating a formalized, idealized phonemic or segmental stage, it should no longer be taken for granted that speech recognition is a normalizing, abstracting process. Variation in the speech signal may be as important to our understanding and encoding of speech as the regularity of the phonemes and the words they comprise.

Introduction: The Speech Signal

Although usually recognized only as an acoustic phenomenon, the speech signal is actually a dynamic complex of interacting sources of information transmitted through multiple sensory modalities (Summerfield, 1983). Sign language and lip-reading provide examples of alternate modalities that can encode and express language. While these communication modes are usually associated with the deaf who cannot otherwise perceive spoken speech, research has found that lip-reading is also a basic part of normal-hearing listeners’ speech perception processes. So basic is lip-reading, in fact, that Sumby and Pollack (1954) found that viewing the face of the talker in a speech-in-noise identification task was equivalent, for increasing accuracy, to a 15dB increase in the signal-to-noise ratio for increasing identification accuracy. McGurk and MacDonald (1976) reported that phonemic perception relied equally on visual and auditory information when subjects were asked to identify phonemes from an audio-video. The sense of touch can even be utilized in speech communication with a glottal pulse train that signals the presence or absence of voicing during articulation (Summerfield, 1987). Deaf listeners who place their hands over the larynx of talkers exploit the same principle. Clearly, then, the speech signal can be and is transmitted through multiple means even for normal-hearing listeners. But exactly what kind of information is being transmitted?

Speech perception is usually considered to be the process by which listeners turn speech into a meaningful string of words and ideas (Goldinger, Pisoni, & Luce, 1996). Yet there is far more to be found in the speech signal than simply a sequence of words or phrases. Information about the age, gender, and the emotional state of the talker as well as information about the talker’s background and socioeconomic status is carried in the speech signal. Recent investigations have found that listeners do extract and encode this type of information (Goldinger et al., 1996; Labov, 1963). The traditional view of speech
communication partitions the information carried in the speech signal into two parallel streams: linguistic information such as phonemes, words, and phrases and extra-linguistic or “indexical” information such as physical or emotional state, gender, dialect, etc (Ladefoged & Broadbent, 1957; Pisoni, 1993). This division is based on the meta-theoretical notion that speech perception operates only on linguistic information and hence extra-linguistic information does not contribute to the basic stages of speech processing. Of course, knowledge about the talker’s gender and emotional state could contribute to our understanding of speech, but the contribution of this information has traditionally been viewed as part of the context that contributes to our understanding after speech processing has occurred (Ladefoged & Broadbent, 1957; Abercrombie, 1967).

The linguistic versus extra-linguistic division is problematic mainly because the information in the speech signal does not seem to fall naturally into this dichotomy at all. While the emotional state or age of the talker may have little affect on the listeners’ phonemic perception, it is obvious from everyday experience that regional dialect and accent can greatly affect the acoustic realization of speech sounds. Moreover, the talker’s age and gender also dramatically affect the acoustic realization of phonemes (Peterson & Barney, 1952; Liljencrants & Lindblom, 1972). The effects of such extra-linguistic factors as gender and age have important consequences for the realization of linguistic information in the speech signal and thus may not be so extra-linguistic after all. Thus, it appears that the linguistic vs. extra-linguistic distinction is an expedient assumption of speech processing and language theories rather than reflecting an actual dichotomy in the speech processing system itself.

Processing the Speech Signal

We can understand the motivation behind the linguistic vs. extra-linguistic division of speech signal information by considering the assumptions of traditional models of speech processing. From its beginnings, speech research was heavily influenced by formal linguistics and hence viewed speech perception as the process that recovers phonemes and words from the speech signal (Halle, 1956; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Pisoni & Lively, 1995). The notion of spoken language as simply an acoustic realization of written language had major consequences for speech research. In particular, researchers looked for the counterparts to the serially ordered, formal and discrete letters of written speech (Licklider, 1952; Halle 1956). Even when it became clear that spoken language was anything but a serial and discrete transmission of information, speech researchers maintained the notion that the process of speech perception recovered and identified formal linguistic units such as phonemes or segments (Liberman et al., 1957; Liljencrants & Lindblom, 1972). Thus, the speech signal was conveniently divided up into two sources of information: abstract linguistic information which presumably contributed to the basic process of phoneme identification and word recognition and extra-linguistic information such as the talker’s gender or accent which was simply filtered out from the speech signal during initial processing (Halle, 1956; Brown 1990).

The notion of variable, extra-linguistic information as noise which is removed from the speech signal underlies both traditional feature-based phoneme recognition models as well as contemporary kinematic models which view speech perception from an ecological standpoint (Pisoni & Lively, 1995; Klatt 1989; Gaver, 1993; Fowler, 1986). According to both of these approaches, the speech signal undergoes a series of processes in order to extract the symbolic phoneme, segment, or articulation movement (Studdert-Kennedy, 1970). The final result is a translation of the speech signal into an abstract, symbolic idealized string of formal linguistic units which are devoid of such information as talker gender or age (Brown, 1990). Again, these models do not necessarily require that extra-linguistic information be removed altogether from the cognitive system of the listener, but rather, this information is processed by non-linguistic systems.
Encoding the Speech Signal

The result of the traditional view of speech processing is an abstracted, formal representation of speech. It is no surprise, then, that the memory structures these views posit for the encoding and storage of speech lack both variability and extra-linguistic content. From the first conceptualization of the mental lexicon as mental dictionary, it was assumed that the memory representation of speech was formal, abstract and generic in nature (Oldfield, 1966). The traditional models fit well with the then accepted prototype or abstractionist view of memory as a storage of idealized memory components. Hence the mental lexicon and memory for other linguistic units were presumed to be prototypical representations of abstract, formal linguistic units. Again, however, no researcher claimed that gender, age or talker identity had no effect on the listener – obviously we can recognize voices and gender - it was simply ignored in their models of speech perception. Thus, whereas our memories for people, places, and events utilize a rich mental representation, our memories for speech events are a simplified, abstracted idealized representation divorced from the age, gender, or emotional state of the persons who spoke.

After traditional accounts of speech processing and storage are employed, we are left with a lexicon which is little more than Webster’s Dictionary encoded neurally – with similarly abstracted and formalized organization (Oldfield, 1966). However, a growing body of new research casts doubt on both the prototypical model of the mental lexicon as well as the formal linguistic notions that have influenced speech-processing research (Pisoni, 1990).

Measuring Speech Perception and Encoding

Challenges to the notion of filtering, abstractionist speech processing were uncovered, but unrecognized, early in the history of speech research. As early as 1955, Peters found that identification of words suffered when multiple talkers spoke the stimulus lists, even though individually the talkers could be easily understood (Peters, 1955). In 1957, Creelman also found that identification of words in noise suffered when stimulus lists were spoken by multiple talkers (Creelman, 1957). These findings were subsequently replicated by Mullennix, Pisoni, and Martin (1989). If the basic process of word recognition begins by filtering the speech signal of extra-linguistic information like talker identity, gender and dialect, we are unable to explain this finding.

Vowel classification experiments found that speakers were able to accurately classify vowels produced by children even when the acoustic characteristics of these vowels were highly dissimilar to adult productions (Gerstman, 1968). Again, unless the extra-linguistic knowledge of a talker’s age is somehow utilized during speech perception, it would seem that listeners would be unable to understand the vowels produced by children unless they used qualitatively different criteria for acoustic-phonemic categorization. Though a robust and repeated finding in early speech research (Miller, 1951), the effects of different forms of variability on speech processing were largely precluded from subsequent experiments due to their basic design. The complexity and cost of stimulus creation as well as the general notion that variability was merely a nuisance factor resulted in most experiments utilizing small stimulus sets and even smaller talker pools. Thus, most of speech research until the last two decades was quite simply unable to assess the effects of stimulus variability and this trend had major effects on research and theory during this time.

As the small but important lineage of speech variability research became more widely known, researchers began designing experiments that explicitly addressed these topics. Kuhl and Miller (1982) found that pre-linguistic infants could accurately discriminate the vowels produced by three different talkers. It appears, then, that even infants are able to process and utilize the variability inherent in different talkers’ speech. In word recognition paradigms it was found that, like Creelman’s (1957) results,
multiple talkers hurt performance (Summerfield & Haggard, 1983). Again this result runs counter to what would be expected from an abstractionist perspective.

The last decade has seen a dramatic increase in speech research specifically designed to measure the effects of stimulus variability. In a serial recall task, Goldinger et al. (1991) found that presentation rate interacted with the number of talkers in the stimulus ensemble to affect subject performance. At slower presentation rates, subjects were better at recalling lists of words spoken by multiple talkers. Multiple talkers hurt performance when presentation rate was faster, however. These results seem to indicate that variable voice information was utilized at slower presentation rates to help recall and thus must have been present in the early neural representation of the words. Using a continuous recall memory task, Palmeri, Goldinger, and Pisoni (1993) found that spoken words were better recognized at test if they were repeated by the same talkers as the original presentation. Moreover, Palmeri et al. also found that performance was worst for words repeated in a new voice from a different gender. Palmeri et al.'s experiment is also noteworthy because subjects were never explicitly instructed to attend to gender or talker identity during the study phase in any of the experiments. These results are simply unexplainable in a framework of abstractionist speech processing which posits that extra-linguistic information such as a talker’s identity and gender are stripped from the speech signal before encoding.

In a perceptual identification experiment, Goldinger (1992) found that implicit memory for talker voice attributes was retained in memory long after perceptual analysis was complete. Subjects returned for follow-up tests 5 minutes, one day and one week after initial word identification trials. As long as one week after the original test, subjects showed accuracy improvements for word repetitions in the same voice as they heard during initial testing and accuracy decrements for difference-voice repetitions. Moreover, these effects were not significantly different from the effects observed in the original testing sessions.

Speaking rate has also been found to influence recall in a manner that is similar to talker identity. Nygaard, Sommers, and Pisoni (1992b) found that recall was better for word lists spoken at the same rate at test as at study. In perceptual learning experiments, talker variability has been found to have a beneficial affect on acquiring and retaining phonetic contrast perception for Japanese listeners (Logan, Lively & Pisoni, 1990; Lively, Pisoni, Yamada, Tohkura & Yamada, 1992). Lively et al. (1992), in particular, found that listeners could better retain their perception of the English /l/ and /r/ contrast when they were exposed to a large corpus of stimuli spoken by many talkers.

Even more recently, Remez, Fellowes, and Rubin (1997) found that subject performance in a sinewave sentence identification task indicates that both talker and phoneme identification rely on the same basic processes. In one condition of their study, subjects listened to naturally produced sentences and then had to judge which of two synthetic sinewave sentences had been produced from the original sentence. In another condition, subjects listened to sinewave processed sentences spoken by familiar voices and were asked to identify the talker. The pattern of results across these and another conditions led Remez et al. to conclude that subjects used the same information in processing sinewave sentences to recover both phonemes and talker identity. Thus the distinction between linguistic and extra-linguistic information does not clearly hold for speech recognition processes as basic as phonemic perception.

In summary, we find from a series of studies that variability has important effects for speech perception that cannot be explained from a traditional abstractionist view of speech processing. Remez et al.'s (1997) research is particularly important because it indicates that this variability plays an important role in the earliest stages of speech processing. In addition, talker variability can be encoded in both an incidental and implicit way with neither the awareness nor the intention of the listener (Peters, 1955; Creelman, 1957; Goldinger, Pisoni & Logan, 1991; Palmeri et al., 1993). For such information to be
encoded in memory would require a memory structure for speech events much more rich in information that the abstract, idealized, symbolic representations usually described. These results also require a speech recognition process which, rather than filtering variability from the signal processes, utilizes, and encodes signal variability. The present experiments were designed to further explore the nature of variability in speech processing of sentences.

**Sentence Processing**

While the role of variability in phoneme and word processing has been examined recently by researchers, little research has been done with spoken sentences. Rather than being the result of a purposeful avoidance of sentence processing experiments, this situation is more likely a result of the assumption that sentence recognition is simply the concatenation of word recognition processes. It is clear from an acoustic analysis of speech, however, that isolated words and words in sentences show significant differences in physical realization (Klatt, 1986; Pisoni & Luce, 1987). In a gating study by Salasso and Pisoni (1985) they found that words removed from their acoustic sentence context were accurately identified only 50% of the time. Clearly, there are important differences between isolated and spoken words that may have implications for the role of variability in sentence processing. If sentences are encoded or processed differently from isolated words, then the effects observed in isolated word research may not be repeated. Due to their length and increased information content and presentation rate, sentences may also require different strategies in processing which may obscure the effects of stimulus variability. These questions concerning the similarities between isolated word and sentence recognition processes can only be addressed by sentence-based research.

When sentences have been used in speech perception experiments in the past, much as the case with single word or phoneme stimuli, few talkers recorded the sentences and the corpus itself was very small. Still, however, several important experiments have utilized sentences as stimuli with the research of Karl (1996) being one of the main motivations behind the present study.

Geiselman and Bellezza (1976) found that talker gender was incidentally encoded in memory for sentences in a recall task. In a similar set of experiments, Fisher and Cuervo (1983) examined memory for extra-linguistic attributes of sentences as they related to sentence comprehension. In their study, subjects were more likely to remember the gender of the speaker or the language of the sentence when this information was important to sentence comprehension.

While the earlier research of Geiselman and Fisher did uncover some effects of stimulus variability, a series of recall experiments conducted by Karl found no effects for talker variability (Karl & Pisoni, 1994; Karl, 1996). Across several variations on a basic free recall task in which subjects heard and then transcribed lists of sentences spoken by one or more talkers, Karl found no significant effects for talker variability on recall performance. In the experiment, sentences were presented in blocks spoken by multiple or single talkers and subjects were asked to recall the sentences by transcribing them in any order when a tone sounded after each block. Karl found that recall was not significantly affected by the number of talkers within each block. Karl concluded that the nature of a free recall task that emphasized rehearsal might not be suitable for uncovering effects of surface features such as talker identity. Karl’s suggestion was the elaboration and extraction that his experiment may have encouraged in the subject obscured the surface information present in the stimuli. Thus we find a mixed message from previous sentence processing research examining the role of variability in speech processing. We also find motivation for the present study.
Recognition Memory

Considering Karl’s conclusions that a serial recall paradigm is inappropriate for research involving talker-specific variability, we selected a discrete recognition memory paradigm for these experiments. The recognition memory design was also chosen because of the large body of isolated word and speech segment research carried out in this paradigm (Egan, 1948; Snodgrass & McClure, 1975; Squire, Shimamura & Graf, 1985; Henson, Rugg, Shallice, Josephs & Dolan, 1999).

In a discrete recognition memory experiment, the procedure is divided into two parts: a study phase and a test phase. During the study phase, subjects review a collection of stimuli and may additionally perform some specified task after stimulus presentation. Depending on the goal of the experiment, the study phase task may be as simple as controlling the pace of stimulus presentation or the task may require additional judgments or actions with regard to the stimuli. Though the stimuli in these and previous speech experiments are auditory speech samples, the recognition memory design can accommodate principally any set of stimuli such as pictures, objects, or visual speech (Kim, Andreasen, O'Leary, Wiser, Ponto, Boles, Watkins & Hichwa, 1999; Cornell, 1980).

Once the study phase has been completed subjects may perform an interpolated task before the test phase begins. Again, this task can be a distractor task to prevent rehearsal or a task designed to clear short-term memory. No interim task was performed in this series of experiments so subjects moved directly to the test phase.

The test phase, much like the study phase, consists of subjects reviewing a new collection of stimuli. The stimuli used during the test phase consist of some or all of the study phase stimuli mixed with additional new stimuli. The subjects’ task in this phase of the experiment is to make a recognition judgment about the stimuli such as whether this stimulus was presented during the study phase (old stimuli) or not (new stimuli). The test phase stimuli are designed to resemble study phase stimuli so that performance in the test phase gives a measure of the discriminability of the new stimuli from the old stimuli as well as a general measure of memory for study phase items. In the present series of experiments signal-detection analyses were performed on response data.

Signal Detection Analysis

The responses to test phase stimuli in these experiments take the form of old (present at study) or new (not preset at study) judgments. Despite the fact that only two responses are possible, there are actually four different response types for these stimuli – two old and two new. Signal detection theory allows us to analyze and make sense of these four response types. The two old response categories are “hits” – correctly labeling an old sentence as old and “false alarms” – incorrectly labeling a new sentence as old. The two new response types correspond to correct new judgments applied to new sentences – “correct rejections” and incorrect new judgments applied to old sentences – misses.

Whereas a standard analysis of performance based on accuracy would lump both old judgments and both new judgments together, a signal detection analysis allows more information to be recovered from the response data. In addition to the scoring for hits, misses, false alarms, and correct rejections, two additional measures can be derived under signal detection analysis. $D^\prime$ is one of these measures computed by subtracting the false alarm rate from the hit rate. The function of $D^\prime$ is used to assess how well subjects can discriminate truly old items from new items. Without the additional insight $D^\prime$ can provide, a subject could respond old to all sentences and appear to have perfect accuracy for judging old sentences – even when the subject may not have been able to identify old sentences at all. Thus, $D^\prime$ gives
an indication not just of the percent correct, but also of the overall trend of accurate and inaccurate judgments and allows a better assessment of the effects on recognition performance.

A second derived measure is called beta (B) and is computed by normalizing the score for correct rejections. Beta is a measure of the underlying tendency for a given subject to make a specific response. Thus, a subject who responds old to all sentences would have a high bias for old responses or, conversely, would be very bad at making new judgments. These two independent measures provide important information about the underlying sensory and decision processes used by subjects during the experiment. For the current experiments the “signal” to be detected is the oldness or newness of a sentence. As the following discussion will show, signal detection analysis allows us to extract important information from an otherwise simple pattern of old and new responses in test phase data.

**Experiment 1: Transcription at Study**

Experiment 1 represents the prototype for the four experiments described in this paper. Both the corpus of stimuli and the materials described in depth here are identical across all four experiments.

**Method**

**Subjects**

Subjects were 33 undergraduate psychology students at Indiana University who were given class credit in an introductory course for their participation. Requirements listed on the sign-up sheet for subject participation specified native English speakers with normal or corrected vision and average typing ability. These requirements were not formally enforced - though each subject was informally assessed by the experimenter before participation. It should be noted, however, that subjects did complete individual data forms which corroborate the assumption that subjects followed the sign-up sheet guidelines.

**Materials**

All stimuli used in this experiment were taken from the Indiana Multi-Talker Speech Database (IMTSD). The IMTSD is a corpus of 100 Harvard sentences (Egan, 1948) recorded by 10 male and 10 female talkers for a total of 2,000 sentence tokens. Each sentence type consists of 5 content words and a variable number of function words. All sentences are meaningful, declarative or imperative statements such as, “These days a leg of chicken is a rare dish” or “Throw the box beside the parked truck.” The Harvard sentences have a long history in speech research and so were used here despite their somewhat dated constructions. Though at times quaint in their grammar, the 100 sentences are easily understood and no subject reported a problem in comprehending a sentence. All of these stimuli were stored in the form of digital computer sound files sampled at 20Hz with 16-bit resolution. For a more complete discussion of the collection and creation of the IMTSD see Karl (1996) and Karl and Pisoni (1994).

By using a collection of descriptive statistics compiled for the IMTSD by Bradlow, Torretta, and Pisoni (1995), the original 100 types were sorted into 80 sentences representing the top 80 most intelligible sentences averaged across all 20 talkers. These statistics were used in the selection process because individual talkers differed in their average intelligibility and thus the same sentences did not fill the top 80% for each talker. The twenty talkers varied in their intelligibility from 81.1% to 93.4% – based on average words correct per sentence across 200 listeners (see Bradlow, Torretta & Pisoni, 1995). While actual population estimates for the range of talker variability have never been formally assessed and hence we cannot relate the magnitude of this difference to that outside the laboratory, it is clear that a considerable amount of variation is present across these 20 talkers in the database.
A computer program was then created to select the individual stimulus files from this collection of 1,600 sentence tokens. The program generated lists of stimulus file names that consisted of a sentence label, a talker label and a gender label. The program worked first by randomly selecting 5 male and 5 female talkers from the list of 20 talkers. The ten selected talkers were the alpha talkers for each stimulus file and the 10 unselected talkers were the beta talkers. From each of the alpha talkers, 4 sentences were then randomly selected from the 80 sentence types. The program additionally monitored sentence selection within each set so that 40 different sentence types were chosen for each file. These sets of 40 sentences were used for the study phase stimuli. Stimulus files were randomly sampled after generation to check that the program had selected a correct set of sentences.

The test phase stimuli were created by processing the study phase files arrays. First, the second and fourth sentences of each alpha talker were duplicated and their talker label changed to a beta talker. Each of these 20 study phase sentences was then replaced by a randomly selected new sentence type spoken by a beta talker. The first and third sentence of each alpha talker were likewise duplicated but with their sentence label changed to a randomly selected new sentence. The result of these operations was a set of test stimuli consisting of 80 sentences matched to the study stimuli.

The following notation will be used to identify the sentences in the test phase sets. A sentence spoken by an alpha talker that appears in both the study and test phase files will be an O_{O}O_t (old sentence/old talker) sentence. New sentence types spoken by alpha talkers at test will be N_{O}O_t (new sentence/old talker) sentences. A sentence type spoken by a beta talker that appears in both the study and test sets is an O_{N}N_t (old sentence/old talker) sentence. Finally, N_{N}N_t refers to a new sentence type spoken by a beta talker (new sentence/new talker). Each test phase file consists of 20 sentences each of OO, NO, ON, NN.

These elaborate stimulus selection strategies were undertaken to avoid any confounds from the inherent differences in sentence and talker intelligibility. It is important that the differences between items inherent in any corpus of speech samples be minimized through randomization. While an item-by-item analysis of each stimulus set file was not undertaken, the multi-step randomization procedure could reliably be expected to evenly distribute the effects of both talker and sentence differences.

Stimulus presentation and data collection were controlled by two computer programs which ran the study and test phases. Sentences in both phases were presented to subjects binaurally over matched Beyer Dynamic DT-100 headphones at a comfortable listening level (75 dB SPL). The programs ran on multiple personal computers and data were written to formatted text files. The keyboard and electronic button-boxes were used to collect responses.

Procedure

Upon completion of consent and credit forms as well as subject information sheets, each subject took a seat in a computer kiosk. Subjects were instructed to find a comfortable position from which to view the CRT monitor and reach the computer keyboard. Subjects were handed headphones and then given instruction as to the task they were to perform. Subjects were told they would be hearing spoken sentences over their headphones. After each sentence had ended, subjects were asked to type in what they heard at an arrow input prompt that appeared on the monitor. Subjects were told the sentences would be simple, meaningful statements in English. Finally, subjects were told that phonetic spellings were acceptable and that blanks could be left for words they could not understand. Most importantly, subjects were not told anything about the ensuing memory test nor were they instructed to attend to any particular
property or set of attributes of the sentences. Subjects were allowed to ask questions to clarify any unclear instructions and then each subject began the experiment.

Subjects hit enter on the keyboard to begin the study phase of the experiment. Upon hitting enter, the program started by reading in a specified study stimulus set and randomizing the files before presentation. A one second delay preceded the beginning of each sentence. Once the sentence had played in full, an arrow prompt -> appeared which was the signal for subjects to type in the sentence they heard. After typing the sentences and making any corrections, subjects hit enter to record their response and advance to the next sentence. Upon completion of the last transcription, the program ended and subjects waited for the experimenter to initiate the test phase of the experiment. Despite the self-paced nature of stimulus presentation, all subjects usually finished the study phase about the same time within 10 to 15 minutes.

Subjects were again instructed before the beginning of the test phase. Subjects were told that in this phase of the experiment they would be hearing 80 total sentences. Subjects were then shown the button boxes used for their response input. The button box consisted of two buttons each with a corresponding LED indicator. Subjects were instructed to judge each sentence as “old”, meaning it was heard during the study phase – or “new”, meaning it was not heard during study. Each button on the button box was labeled as either “old” (1) or “new” (2). Both the number and word label for the button choice appeared on the button box itself as well as the on screen instructions that began each experiment. Subjects were told to guess “old” or “new” in the event that they could not otherwise judge a sentence. Subjects were given no explicit criterion for judging a sentence as “old” or “new” beyond the instructions that the sentence “appeared in the study phase” description. Several subjects did report during debriefing, however, that one of the determining factors of their judgment was a particular speaker’s distinctive way of pronouncing certain sentences that were presented during the study phase.

Subjects began the test phase by hitting enter after which a brief “Ready” message flashed onscreen before the sentence began. After each sentence was presented, the familiar arrow prompt appeared at which point subjects recorded their judgment. Subjects could be certain the computer had received their response by the light above the pushed button flashing briefly and also by the flash of a new “Ready” message onscreen. The minimal response required of the subjects in this phase of the experiment usually resulted in finishing times of less than 10 minutes despite there being twice the number of sentences as in the study phase. Like the study phase, all subjects finished around the same time despite the self-paced stimulus presentation. Upon completion of the test phase, subjects were debriefed and asked for any comments concerning the experiment and then allowed to leave the laboratory.

The data generated for each phase of the experiment were recorded in two corresponding data files. Each file contained a record of the stimulus set files used as well as the subject and session number. Each study phase file also contained the subject-transcribed sentences paired with the actual content of each sentence. The test file sentences contained a numeric code corresponding to the old (1) or new (2) judgment of each sentence paired with the actual label for each sentence.

For ease of scoring, a computer recovery program was written to process each test-phase file into a formatted data file. The program analyzed the subjects’ responses and returned tallies of the number of correct and incorrect judgments for each of the four test sentence categories. Correct judgments for OO and ON sentences were ‘old’ or ‘1’ while ‘new’ or ‘2’ were the correct judgments for NN and NO sentences. These resulting files were then analyzed using Microsoft Excel.
**Results**

Due to a computer malfunction that corrupted 8 subject data files as well as additional computer problems that interrupted 3 students during the experiment, 11 subjects’ data were discarded before analysis. Finally, an additional subject was eliminated from analysis based on his negative d’ score (see below). Thus the data from 21 subjects was used in analysis.

Study phase files were first analyzed to measure transcription accuracy. Following procedure common in transcription task scoring (Karl, 1996), phonetic spellings and obvious typos were counted as correct transcriptions. This was clearly an easy task. All subjects scored above 85% of words accurate. Transcription errors were usually systematic such as ‘cow’ instead of ‘cop’ or article substitutions like ‘a’ for ‘the’. From this we can conclude that when subjects did not fully hear parts of a sentence they attempted to compensate using partial stimulus information. The use of partial information, termed “sophisticated guessing,” has both a long experimental and theoretical history in psycholinguistic research (Solomon & Postman, 1952; Savin, 1963).

**Recognition Accuracy:** Figure 1 summarizes recognition accuracy results. The boxes containing probability values span columns whose difference is significant at that probability value.

![Figure 1: Recognition Accuracy for Experiment 1](image)

*Figure 1: Recognition accuracy for Experiment 1 grouped by sentence category. Probability boxes span the columns whose difference is significant with that probability in paired t-tests.*

Test phase responses were first tallied into correct judgment totals for each of the four sentence categories. For analysis, these totals were expressed as probability of correctly recognizing a sentence. Thus, a subject judging OO sentences as old 17 out of 20 times would have 85% accuracy. Subjects generally showed better accuracy in judging OO sentence than ON sentences. Individually, only 3 of the 21 subjects had higher accuracy for ON sentences. Overall, performance accuracy for OO sentences
ranged from 60% to 100%, ON sentence accuracy varied across a smaller range from 70% to 95% correct.

The average percent correct for OO sentences was 89.5% while average correct for ON sentences was 84.5%. This difference was significant using a paired t-test, $t(20)=2.334$, $p=0.027$. No significant differences in accuracy were observed for NN sentences and NO sentences although performance on both these sentence types was near ceiling with accuracy averages of 99% for both NN and NO sentences. The ceiling effect is further revealed by the range of accuracy scores that varied only from 95% to 100%.

Two means were computed for each subject from the category pairs OO/ON and NN/NO. One average thus represents the old sentence average accuracy (OAA) while the second reflects new sentence average accuracy (NAA). Average OAA was 87.5% while average NAA was 98.4%. This difference was significant with a paired t-test, $t(20)=4.96$, $p<0.0005$. This suggests that new sentences are easier to judge than old sentences and that subjects display a response bias to response “new” more often than “old.”

A 2x2 ANOVA was then performed on the accuracy results. This analysis yielded a significant main effect for sentence type, $F(3, 80)=13.9$ and $p<0.0005$. Neither the main effect for voice nor the 2-way interactions were significant. The results ANOVA further corroborate findings from the t-tests that old sentences were recognized significantly better than the new sentences. The observed ceiling effect, however, makes these conclusions tentative at best.

d’ scores: Two d’ measures for each subject were then computed from the recognition data. Hits were scored for every judgment of ‘old’ given to an OO or ON sentence. Likewise, misses were scored for ‘new’ judgments of these sentences. A correct rejection was scored for each ‘new’ judgment given an NN or NO sentence. Finally, false alarms were scored when a subject judged either of these two types as ‘old.’ A d’ measure of sentence discriminability for sentences produced by “old” voices was calculated by subtracting the average false alarms for NO and NN sentences from hits for OO sentences. Subtracting average false alarms for NO and NN sentences from hits for ON sentences generated the d’ measure for new-voiced sentences. The false alarm rates for NO and NN were averaged because NN and NO accuracy were not significantly different. This measure provides a better overall measure of subjects’ false alarm rate. After the d’s had been calculated in this manner for all subjects, one outlier emerged had negative d’. Since the reasons behind this discrepant performance were unknown, the subject was removed from the final analysis and thus all results discussed here exclude this subject. The average d’ was 3.56 for old voices while the average d’ for new voices was 3.24. The paired t-test for the difference in d’s yielded a significant difference, $t(20)=2.522$, $p=0.02$ and thus subjects were able to discriminate between “old” and “new” sentences better when the sentences were in an old voice at test that when the sentences were in a new voice.

Discussion

The results of this experiment suggest two important—and related—conclusions. First, these results provide support for the proposal that the memory representation for spoken sentences is rich enough to preserve voice information. Clearly, the significantly higher accuracy in OO identifications compared to ON identifications suggests that voice information was preserved in memory along with the lexical organization of the sentence itself. Thus, we can conclude that in this task, voice information was responsible for as much as a 5% improvement in recognition accuracy although performance was at or near ceiling. Moreover, under the assumption that subjects were not explicitly attending to voice information, the processing and storage of this information appears to be automatic without explicit conscious attention or awareness. The assumption about subject behavior seems plausible, since voice information, much less a memory task, was not mentioned in subjects’ instructions. Still, it is possible
that a confounding number of subjects did specifically attend to voice information during the study phase. A more extensive post-experiment interview than the one conducted for this experiment may be able to resolve this issue.

Secondly, as an obvious corollary to the first conclusion, these results suggest that voice information can also be utilized to aid performance in a recognition memory paradigm. A rich memory representation would be useless without search and retrieval processes that can make use of that additional information. Further, as many subjects did report using voice information in at least some of their judgments, this experiment cannot speak as to the automaticity of voice information use in the recognition process.

The \( d' \) results show that subjects were better able to discriminate “new” sentences from “old” sentences when the sentence was spoken by an “old” talker. This provides further evidence for the encoding of voice information since there should be no difference between new voice and old voice sentence discriminability were voice information absent from memory. The fact that the difference in \( d' \) occurred without subjects being instructed to attend to voice information further establishes that voice information encoding is an automatic feature of speech event encoding.

Finally, we can see from the data that, particularly in the NN and NO category, performance was near ceiling. This ceiling effect obscures important statistical differences. On the one hand, the accuracy improvement of voice information could be higher than the average 5% here recorded. Additionally, the difference in accuracy between NN and NO sentences may simply have been too small to detect in a collection of scores whose average was only 1% below ceiling. The following experiment was designed to lower subject performance from the ceiling and hence give a better estimate of voice effects in sentence recognition memory.

**Experiment 2: No Transcription**

In an attempt to lower subject accuracy and move performance down from the ceiling, we modified the study task used in Experiment 1. It is very likely that the depth of processing encouraged by the sentence transcription task resulted in a memory representation that was based on conceptual processing which made the memory particularly accurate for the lexical sentence information and hence caused the ceiling effects. By removing the transcription task from the study phase, we hoped to bring performance down from ceiling and thus yield a more accurate measure of voice effects in sentence recognition.

**Method**

**Subjects**

Subjects were 21 undergraduate psychology students at Indiana University who were given class credit in an introductory course for their participation. Requirements listed on the sign-up sheet for subject participation specified native English speakers with normal or corrected vision and average typing ability. Again, subjects were not formally assessed for meeting these requirements, but subject data forms do corroborate that subjects did follow these guidelines.

**Materials**

Stimulus materials for this experiment were taken from the same corpus as Experiment 1. The randomly generated stimulus set files used in the previous experiment were also used here. The same
experimental control programs and data-processing routines described for earlier were also used in this experiment.

Procedure

The basic procedure used in this experiment was identical to that described in Experiment 1 except that subjects were not required to transcribe the sentences after each stimulus presentation. Instead, subjects simply hit the enter key on the keyboard to advance to the next stimulus. Subjects were instructed to listen and pay attention to each sentence. As in Experiment 1 no explicit mention was made of voice information in these sentences or the ensuing recognition task.

Results

Recognition Accuracy: Figure 2 summarizes recognition accuracy results. The boxes containing probability values span columns whose difference is significant at that probability value.

For OO sentences, subjects accurately judged 83.3% of sentences on average as ‘old’ in the test phase. ON sentences showed 77.6% average accuracy for “old” sentences with a significant difference of 5.7% in a paired t-test, \( t(40)=2.274, \ p<0.05 \). While performance in NN and NO sentences was successfully lowered from ceiling, the NN average of 89% and NO average of 87% were still high. Like Experiment 1, there was no significant difference between these two scores.

Old sentence average accuracy (OAA) and new sentence average (NAA) accuracy scores were again computed for each subject. Average OAA was 85.9% average NAA was 87.7%. This difference was not significant in a t-test. The results of experiment 1 which suggested that new sentences were easier
to discriminate than old sentences are thus not replicated in this study – though the direction of the difference in average accuracy is the same.

A 2x2 ANOVA performed on accuracy results yielded a significant main effect for sentence type with $F(3, 80)=7.6, p=0.007$. Neither the main effect for voice nor the 2-way interactions were significant.

d’ scores: Two d’ scores for each subject were calculated. One score compared the two old talker categories while the other score compared performance for the two new talker categories. These d’ scores were again calculated using the false-alarm averaging technique described previously. As in Experiment 1, this averaging was justified by a non-significant difference between NN and NO accuracy. The same criterion placed on d’ scores in Experiment 1 was also used here, but as no subject had negative d’s, all subjects were kept in the analysis. The old-voiced d’ average was 2.48 while the new-voiced average was 2.24. This difference was marginally significant, $t(20)=2.01, p<.06$. D’ scores for old voices ranged from 1.14 to 4.2. New voiced d’s ranged from 0.97 to 4.2.

**Discussion**

The results of this experiment show even stronger voice effects in sentence processing than those of Experiment 1. The 5.7% difference between OO and ON sentences was marginally higher than the 5% difference in Experiment 1. While the non-significant difference between NN and NO sentences was replicated in this experiment, scores were successfully pulled from the ceiling. The difference between the NN and NO scores was also slightly larger than in Experiment 1: 1% in Experiment 2 as compared to no difference Experiment 1. The absence of a transcription task had both the desired effect of lowering performance and increasing the difference between old-voiced and new-voiced category pairs.

The lack of difference between accuracy in new sentence judgments for old and new talkers was replicated in this experiment. Since neither of the two experiments so far discussed have found a difference, it appears that subjects can rely exclusively on lexical information to judge a new sentence accurately regardless of talker. When the sentence is old, however, subjects show significant effects for voice information.

We can explain this pattern of results by considering two situations in which a subject is required to discriminate sentences in this procedure. When presented with a new sentence, subjects easily recognize that the subject matter and phrasing of the sentence are novel. Whether the voice is old or new makes no difference since the words themselves are clearly different. When confronted with an old sentence, however, subjects cannot be certain whether the sentence is the same as one heard before or simply similar to previous sentences in subject matter or phrasing. In such problematic cases subjects may then turn to voice information to judge the sentence. Since old voices were the speakers of old sentences and thus subjects could utilize both their lexical and indexical memories for the sentences, and thus it is more likely that subjects will judge OO sentences correctly than ON sentences. In principle, this effect should occur to some extent in new sentence judgments though the ease of these judgments in most circumstances may mask this affect. By making the study task more difficult, a voice affect for new sentences may emerge.

The d’ scores for Experiment 2 were significantly lower than those for Experiment 1. This means that subjects were less sensitive to sentence change after a study task which did not require transcription. This result is not surprising to the extent that transcription requires more attention and conceptual processing than passively listening to a sentence. Thus, subjects would be more sensitive to the semantic differences between sentences after transcription.
It is likely that a passive listening task did not encourage subjects to engage in as intense an attentive process as transcription. Thus, the sentences may not have been as well encoded in memory as suggested by a level of processing perspective (Craik & Lockhart, 1972). This notion is borne-out by the d’ prime data which show a decrease in sensitivity to sentence change from experiment 1. General performance accuracy showing a 6% decline also supports this interpretation. But voice information had more of an affect in this experiment than the previous experiment. Apparently, subjects were able to utilize voice information to aide performance in judgment when their conceptual memory for the lexical/semantic characteristics of the sentences was inadequate. Since there was less conceptual information concerning sentences in Experiment 2 by which subjects could make accurate judgments, we can explain both the decrease in performance and discriminability. Moreover, since subjects have made recourse to voice information to compensate for less conceptual information, we can also explain the increase in voice effects between Experiment 1 and Experiment 2.

While the change in study task between Experiments 1 and 2 had the desired effect in pulling performance away from ceiling, it still did not reveal any significant voice effects for new sentences. Unless we posit that new sentences were somehow processed and judged differently from old sentences, there should also be voice effects in old sentence judgment accuracy. Performance was still very high across all conditions and hence the true magnitude of voice effects may still be partially masked. Experiment 3 was designed to lower performance further and hopefully allow more extensive voice effects to be uncovered.

**Experiment 3: Word List Transcription**

Although the change in study between Experiment 1 and 2 was enough to lower performance from ceiling, subjects still showed extremely accurate judgments for all sentence categories. Thus, in this experiment an additional word list transcription task was added to the study phase to further lower performance and hopefully reveal any voice affects that might be present.

**Method**

**Subjects**

Subjects were 21 undergraduate psychology students at Indiana University. Seventeen of the subjects received credit in an introductory psychology. Four other subjects were recruited from an introductory psychology course and paid $5 for their participation. Requirements listed on the sign-up sheet for subject participation specified native English speakers with normal or corrected vision and average typing ability. Subjects were not formally assessed for meeting these requirements, but subject data forms do corroborate that subjects did follow these guidelines.

**Materials**

Stimulus materials for this experiment were taken from the same corpus used in experiments 1 and 2. The randomly generated stimulus set files used the earlier experiments were also used in this experiment. An additional list of word pairs randomly selected from the content words of 40 of the 80 sentence types was also used during the study task. For each subject, some of the word pairs would be heard in study phase sentences while others would not. In the test phase, all the word pairs would appear in sentences. The word pair list was designed to reduce the conceptual processing of sentences during study and create interference in recall during test and thus encourage subjects to make more use of voice information in their judgments. The same experimental control programs and data-processing routines described for 1A were also used in this experiment.
Procedure

The procedure in this experiment was the same as in Experiment 2 except that subjects were also required to type in word pairs after each sentence presentation during the study phase. The word pairs were numbered 1 to 40 and presented on a printed page placed beside the keyboard. Subjects were instructed to type in the correspondingly numbered word pair after each sentence was presented. Subjects were further instructed to only read or type word pairs after they had heard the complete sentence. This instruction was given to prevent subjects from reading the word list before sentences were finished and not accurately encoding study sentences. As in the previous experiments, no mention was made of voice information or the recognition memory task.

Results

Recognition Accuracy: Figure 3 summarizes recognition accuracy results. The boxes containing probability values span columns whose difference is significant at that probability value.

![Figure 3: Recognition Accuracy for Experiment 3](image)

For OO sentences, subjects accurately judged 70% of sentences as ‘old’ in the test phase. ON sentences showed 60.7% accurate judgments at test for an arithmetic difference of 9.3% which was significant in a paired t-test, \( t(20)=4.35, p<0.001 \). Performance in NN and NO sentences was successfully pulled from ceiling with an NN average of 79% and NO average of 77.5%. However, like experiment 1 & 2, the difference between these two scores was not significant.
Old sentence average accuracy (OAA) and new sentence average (NAA) accuracy scores were again computed for each subject. Average OAA was 70.7% while average NAA was 80.2%. This 9.5% difference neared significance, \(t(20)=2.00, p=0.06\). This difference is again in the same direction as Experiments 1 and 2 and is nearly significant. Thus we find slight support for the conclusion that new sentences are easier to discriminate than old and additional evidence that subjects has an overall bias to respond “new.”

A 2x2 ANOVA performed on accuracy results yielded a significant main effect for sentence type with \(F(9, 80)=14.1, p<0.005\). Neither the main effect for voice nor the 2-way interactions were significant.

**d’ scores:** Two d’ scores for subjects were calculated. One score compared the two old talker categories while the other score compared performance for the two new talker categories. These d’ scores were again calculated using the false-alarm averaging technique described previously. As in Experiments 1 and 2, this averaging was justified by a non-significant difference between NN and NO accuracy. The same criterion placed on d’ scores in experiment 1a was also used here, but as no subject had negative d’s, all subjects were kept in the analysis. The old-voiced d’ average was 1.46 while the new-voiced average was 1.18. This difference was significant, \(t(20)=4.47, p<0.005\). d’ scores for old voices ranged from .38 to 3.0. New voiced d’s ranged from 0.36 to 2.85.

**Discussion**

Again, the results from this experiment support a mental representation of speech events that includes extra-linguistic information such as voice information. Also important is a third replication of voice effects for old-voiced despite subjects reporting that they did not attend specifically to voice information during study. Thus, it appears that voice information is encoded automatically without the intention of the listener. Research by Remez et al. (1997) suggesting that voice information is utilized during the phonetic identification offers an explanation for this automaticity: The low-level processes of phoneme identification require voice data and hence lead to voice information encoding during the initial unconscious stages of speech processing.

The difference between accuracy for OO and ON sentences was the largest yet seen in this series of experiments at 9.3%. Thus, the additional voice information provided by OO sentences allows for as much as a 9.3% gain in accuracy. For new sentences, voice information again seemed to make no difference as the 1.5% difference between NN and NO sentences was not significant.

The d’ scores for this experiment were the lowest yet seen in this series of experiments. We can interpret these low scores as evidence that the interpolated word-pair transcription task reduced the encoding of sentence information that contributed to higher d’ scores in the earlier experiments. We cannot specify, however, whether the word-pair transcription task simply altered the level of processing the sentences were encoded with at study or if the word-pairs interfered with recognition during the study task.

Accuracy in the new sentence categories was pulled well away from ceiling with averages for NN and NO of 79% and 77.5% respectively. Since these scores are even lower than the old sentence accuracy scores from Experiment 1 – which did reveal a significant voice effect – it seems that voice information is simply not utilized to a significant extent when subjects discriminate new sentences. The previous suggestion that voice information may not be utilized because new sentences are easy to discriminate based solely on their novel lexical and semantic content seems well supported by these results.
From Experiments 1 to 3, we find increases in the accuracy difference between old-voiced and new-voiced sentences while overall judgment accuracy and d’s significantly decline. Thus, we must consider how voice information provides a greater gain in sentence accuracy judgments when the accuracy of the judgments themselves decreases overall. One possibility is that the full sentence transcription task used during study in Experiment 1 required more attention than the word pair transcription of Experiment 3. The additional requirement was met at the expense of voice information encoding and thus each stimulus could not be processed as deeply as in Experiment 1. Since not as much voice information was encoded at study in Experiment 1, subjects could not benefit as much from this information when making test phase judgments.

A second possibility also results in less voice information being available after sentence transcription, but through a different process. It is possible that the full sentence transcription task itself interfered with voice information encoding to the extent that subjects likely rehearsed the sentences while typing them. This rehearsal may have encouraged conceptual coding of the sentences and obscured surface feature information such as talker voice. Thus, voice information may have initially been encoded to the same extent as in Experiment 3, but the process of rehearsal during transcription typing may have attenuated this voice information and resulted in a loss of surface perceptual features. The result is that less voice information was available at test in Experiment 1 to assist in sentence judgments. The study task in Experiment 3, which did not require sentence rehearsal, would not have encouraged conceptual processing that would likely have obscured surface perceptual information. Thus, when subjects were required to judge sentences as old or new in the test phase, this surface information would have aided in their judgment of sentences.

So far, we have examined test phase voice effects after study tasks that did not direct explicit attention to voice information. The final experiment here discussed was designed to assess voice affects after a task that directly encouraged processing of voice information at the time of study.

**Experiment 4: Voice Monitoring**

In the first 3 experiments, we measured the indirect effects of voice information when subjects were not instructed or encouraged to pay attention to these properties of the speech signal. The procedure in the present experiment was specifically chosen to encourage subjects to pay explicit attention to voice information without making the real aims of the experiment explicit. This change in task was considered because we had no baseline measurement for the effects of voice information when subjects were explicitly asked to attend to this surface feature. Without information concerning voice effects in an explicit encoding condition, we cannot judge the relative magnitudes of the voice effects from Experiments 1-3. The results of this Experiment 4 can thus be considered a kind of baseline by which to compare the voice effect results from Experiments 1, 2 and 3.

**Method**

**Subjects**

Subjects were 21 undergraduate psychology students at Indiana University. Fifteen of the subjects received credit in an introductory psychology. Six other subjects were recruited from an introductory psychology course and paid $5 for their participation. Requirements listed on the sign-up sheet for subject participation specified native English speakers with normal or corrected vision and average typing ability.
Materials

Stimulus materials for this experiment were taken from the same corpus used in the three previous experiments. The randomly generated stimulus set files used in Experiments 1 and 2 were also used in this experiment. This experiment returns to the simpler design of Experiments 1 and 2 and thus did not utilize a word-pair list. The same experimental control programs and data-processing routines described for Experiment 1 were also used in this experiment.

Procedure

The procedure used in this experiment was the same as Experiments 1 and 2 except that now subjects were asked to explicitly identify the gender of the talker during the study phase. After each sentence was played, subjects indicated the gender of the sentence-talker by typing in male or female and then hitting enter to record their response. Subjects were allowed to abbreviate their responses to ‘m’ or ‘f’. Subjects were also instructed to guess when they were uncertain as to talker gender. By specifically calling attention to voice information in this deliberate way, it is expected that subjects will show increased effects for voice information in this experiment as compared to Experiments 1-3. These increased effects should emerge in larger differences in accuracy between old and new voiced sentences as well as higher d’ scores.

Results

Recognition Accuracy: Figure 4 summarizes recognition accuracy results. The boxes containing probability values span columns whose difference is significant at that probability value.

![Figure 4: Recognition Accuracy for Experiment 4](image)

Figure 4: Recognition accuracy for Experiment 4 grouped by sentence category. Probability boxes span the columns whose difference is significant with that probability.
For OO sentences, subjects accurately judged 72.9% of sentences as ‘old’ in the test phase. ON sentences showed 59% accuracy at test for an arithmetic difference of 13.8% which was significant in a paired t-test, \( t(20)=3.54, p=0.002 \). Performance in NN and NO sentences was well below ceiling with an NN average of 83.4% and NO average of 74%. For the first time in this series of experiments, this new sentence difference across voice was significant, \( t(20)=2.58, p=0.018 \).

Old sentence average accuracy (OAA) and new sentence average (NAA) accuracy scores were again computed for each subject. Average OAA was 66.4% while average NAA was 76.4%. This 10% difference, the second largest encountered so far, was not significant because of the variability in subject performance. This difference is in the same direction as Experiments 1-3. This result provides converging support for the notion that new sentences are easier to judge than old sentences.

A 2x2 ANOVA performed on accuracy results yielded a significant main effect for sentence type with \( F(3, 80)=13.9, p<0.005 \). The main effect for voice was not significant. Likely the result of significant differences between both OO/ON and NN/NO sentences, the 2-way interaction between sentence and voice was significant with \( F(3.80)=11.54, p=.003 \).

**d’ scores:** Two d’ scores were also calculated for each subject. One score compared the two old talker categories while the other score compared performance for the two new talker categories to test for differences in discriminability between old voiced and new voiced sentences. Despite the fact that this experiment did find a significant difference between NO and NN accuracy, the same false alarm averaging technique used in experiments 1-3 was utilized. This was done to keep all d’s derived in this series of experiments the result of the same calculation procedure and hence to add validity to the overall analyses discussed below. The same criterion placed on d’ scores in experiment 1A was also used here, but as no subject had negative d’s, all subjects were kept in the analysis. The old-voiced d’ average was 1.54 while the new-voiced average was 1.15. This difference was significant, \( t(20)=3.50, p=0.002 \). d’ scores for old voices ranged from .91 to 2.57. New voiced d’s ranged from 0.45 to 1.94.

**Discussion**

The recognition results of this fourth experiment are important for several reasons. First, they provide a baseline of voice effects with which to compare the other experiments. Since the focus was on the gender of the speaker of each sentence, subjects were explicitly encouraged to attend to voice information. This manipulation seems to have the desired effect as differences between old and new voiced sentences showed the largest differences in recognition accuracy of all four experiments. Indeed, the difference between OO and ON sentences in this condition is almost 3 times that measured in Experiment 1. The difference between NO and NN sentences is over 7 times larger than the difference in Experiment 1; however, the ceiling effect may explain some of this difference. The additional voice information provided by OO sentences allows for as much as 13.8% gain in accuracy. For the first time in this series of experiments, NN sentences were judged significantly more accurately than NO sentences. The additional “novelty” that a new talker added to NN sentences was worth as much as a 9.3% increase in recognition accuracy. Thus, when subjects were explicitly directed to attend to voice information, they encoded this information better and hence showed greater differences in performance due to the effects of this additional encoding.

Accuracy in the new sentence categories was pulled well away from ceiling with averages for NN and NO of 81.7% and 72.4% respectively. The voice effect found for the new sentences in this experiment is likely a result of the explicit voice monitoring study task that encouraged encoding behavior for sentences that was not present in the earlier experiments. The possibility that this effect was present but obscured by ceiling effects is doubtful because new sentence accuracy scores in this
experiment were not significantly different from new sentence accuracy in Experiment 3 which did not display any voice effects.

The earlier suggestion that voice information is not utilized for new sentence judgment is contradicted by these data. However, new sentences showed only a 9.3% difference in accuracy based on voice compared with the 13.7% difference for old sentences. While voice information was undoubtedly utilized to distinguish NN and NO sentences in this experiment, the effects were still less than those observed for OO and ON sentences. Thus, rather than the previous conclusions of Experiments 1-3 that voice information was not utilized in distinguishing new sentences, we can conclude that voice information was also used in new sentence judgment. Again, this result is not surprising considering the study task of Experiment 4 that encouraged subjects to attend to voice information. Since voice information was likely to be encoded in the sentence memory representations, subjects would more likely have recognized the voice of an old talker speaking a new sentence. This familiarity may have interfered with the subjects’ ability to judge the sentence based on the lexical/semantic content alone. New sentences spoken by new talkers would not have this added familiarity and thus subjects could be more accurate in judging these sentences as “new.”

These results are also important as a fourth replication of voice effects in recognition memory for sentences. Needless to say, an experimental procedure that calls attention to voice information and then finds effects of that information is not particularly striking. However, these results are still important because subjects again demonstrate incidental encoding of talker identity information in their memories for the sentences. Since no subject was told to pay attention to talker identity apart from gender – and no subject was instructed to remember the sentence or voice of each stimulus – we can reliably conclude that this information is encoded even without the intent of the listener.

As in Experiment 2 and 3, we again find that increases in the accuracy difference between old and new voiced sentences are accompanied by lower overall accuracy and lower d’s. Although subjects were less sensitive in this experiment to sentence change when compared to Experiments 1 and 2, the affect of voice information was significantly larger here for both old and new-voiced sentences. Apparently whatever factors caused the decrease in judgment accuracy did not affect the contribution of voice information to sentence judgment. The difference in study tasks is the best explanation for this result. Even though subjects were instructed to listen to all the sentences fully during study, it is possible that subjects attended to stimulus sentences only long enough to determine speaker gender. When confronted with sentences in the test phase, subjects who only processed sentences to uncover talker gender would find it difficult to judge sentences as old or new. Likewise, sensitivity to sentence change would also be less if subjects more or less ignored lexical and semantic information in favor of talker gender information. Thus, preferentially attending to voice quality over lexical information can explain at once the increase in voice effect across both old and new sentences as well as the decrease in overall accuracy and d’. Clearly, then, some way of insuring that subjects listen to sentences fully – without using a transcription-style task – is needed to make confident conclusions from these results.

**Overall Results and Discussion**

**Between Experiments Recognition Accuracy t-Tests:** Recognition judgment accuracy was compared between all four experiments first by pair-wise t-tests as data became available over this series of experiments. The pattern which may become apparent in the data in which both d’ and recognition accuracy display mostly monotonic behavior from Experiment 1 to Experiment 4 is simply a interesting pattern in the data. Since no particular experimental component was varied systematically across Experiments 1 to 4, the trend in the data is merely coincidence.
Independent t-tests were conducted pair-wise between experiments to test the significance of these changes. Figure 5 summarizes these results.

![Figure 5: Mean Accuracy Grouped by Sentence Category](image)

**Figure 5:** Mean recognition accuracy grouped by sentence category. Probability boxes span the columns whose difference is significant with that probability from unpaired t-tests.

The 6% and 7% decline in recognition accuracy from Experiment 1 to Experiment 2 both neared significance in the unpaired t-tests: $t(40)=1.69$, $p=0.098$ for OO recognition accuracy and $t(40)=1.75$, $p=0.087$ for ON recognition accuracy. OO and ON recognition accuracy in Experiment 3 declined an additional 13% and 17% respectively which were both significant decreases from Experiment 2’s results in paired t-tests: $t(20)=2.73$, $p=0.01$ for OO recognition accuracy and $t(20)=3.33$, $p=0.002$ for ON recognition accuracy. Experiment 4’s accuracy data is a mix of further declines and some small increases: OO recognition accuracy increased with respect to Experiment 3 while ON recognition accuracy remained unchanged. Neither of these values were significantly different from Experiment 3 but were significantly different from Experiment 2 by unpaired t-tests: $t(40)=2.20$, $p=0.03$ for OO recognition accuracy, $t(20)=3.22$, $p=0.002$ for ON recognition accuracy.

ON and NN recognition accuracy show similar trends to “old” sentence recognition accuracy. The average decline of ~1% for both “new” sentence categories from Experiment 1 to 2 was significant in paired t-tests: $t(40)=4.46$, $p<0.0005$ for NN recognition accuracy and $t(40)=5.48$, $p<0.0005$ for NO recognition accuracy. Both “new” sentence recognition accuracy scores dropped an additional average of 1.5% from Experiment 2 to Experiment 3. Again, these small declines were significant in unpaired t-tests: $t(40)=2.57$, $p=0.014$ for NO recognition accuracy and $t(40)=2.52$, $p=0.016$ for NN recognition accuracy. The differences in “new” sentence recognition were inconsistent from Experiment 3 to 4, but both were significantly lower than Experiment 2’s recognition accuracy results in unpaired t-tests: $t(40)=3.40$, $p<.005$ for NO recognition accuracy and $t(40)=2.27$, $p<.03$ for NN recognition accuracy.
The difference scores between the two “old” sentence categories and the two “new” sentence categories were computed by subtracting each subject's ON recognition accuracy from their OO accuracy. Figure 6 displays the accuracy difference scores for each experiment.

Figure 6: Recognition Accuracy Difference Scores

The same was done for NO and NN accuracy. These new sets of derived measures were then compared in pair-wise t-tests again done as each additional experiment’s data became available. Though successfully pulled from ceiling, the accuracy differences observed in all four experiments were not significantly different from one another. The difference between OO and ON sentence recognition accuracy in Experiment 1 was 4.8% and not significantly different from the 5.7% in Experiment 2. The “old” sentence accuracy difference of 9.3% was significantly different from Experiment 1 in, \( t(40)=4.5,\ p<0.005 \) but this score was not significantly different from either Experiment 1’s or Experiment 4’s “old” accuracy difference. Finally, Experiment 4’s accuracy difference for “old” sentences was 15% and significantly different from both Experiment 2 and Experiment 1 in unpaired t-tests: \( t(40)=2.31,\ p<.05 \) when compared to Experiment 1 and \( t(40)=2.04,\ p=0.048 \) when compared to Experiment 2.

From the pattern of “old” and “new” sentence recognition accuracy, it appears that recognition was easiest after a sentence transcription task and hardest after an interpolated word list or voice-monitoring task. Moreover, the effect of voice on recognition accuracy became larger relative to overall...
recognition accuracy across the four experiments. That is, as average accuracy decreased from Experiments 1 to 4, the voice difference between old sentences increased monotonically. This pattern of results could serve as a guide for future research. Transcription study tasks yield superior recognition accuracy and show comparatively lower voice effects at test than voice monitoring study tasks. In particular, it is clear that a transcription task so thoroughly acquaints subjects with the lexical and semantic content of sentences that “new” sentence judgments are at ceiling. For experiments such as the present series that may look for effects in “new” sentence recognition accuracy, a transcription study task should not be utilized. Likewise, an experimenter looking to minimize voice effects in recognition performance would not want to choose a task such as the interpolated word-list transcription that shows large and significant effects for voice in “old” sentence recognition accuracy.

Old sentence average accuracy scores (OAA) from Experiment 1 were then compared with Experiment 2 OAA in an unpaired t-test that did not reach significance – the difference between averages being only .016%. Old sentence judgments appear to have been equally easy in both experiments. New sentence average accuracy scores (NAA) from Experiments 1 and 2 were significantly different with in an unpaired t-test, \( t(40)=3.72, p<0.001 \) from an average arithmetic difference of 11%. New sentence judgments were apparently easier for subjects in Experiment 1 than in Experiment 2 and easier than “old” sentence judgments in both experiments.

OAA scores from Experiment 2 were then compared with Experiment 3’s OAA in an unpaired t-test which found a significant difference between the two, \( t(40)=2.73, p<0.01 \) with an arithmetic difference of averages equal to 15%. Old sentence judgments appear to have been easier in experiment 1B – not surprising considering the difference in task requirements for the two experiments. The difference in NAA scores between Experiments 2 and 3 was close to significant in an unpaired t-test, \( t(40)=1.84, p=0.07 \) with an arithmetic difference of averages equal to 7.5%. New sentence judgments seem to have been slightly easier in Experiment 2 than in Experiment 3.

OAA scores from Experiment 2 were then compared with Experiment 4’s OAA in an unpaired t-test which found a significant difference, \( t(40)=3.45, p=0.001 \) with arithmetic difference average of 15%. Old sentence judgments appear to have been easier in experiment 1B – not surprising considering the difference in task requirements for the two experiments. The 11.5% difference in NAA scores between 1B and 1D was significant, \( t(40)=2.52, p=0.016 \). New sentence judgments seem to have been slightly easier in experiment 1B than in 1C. The overall observed trend for subjects to respond “new” also contributed to the consistently higher new sentence accuracy – but these judgments were also quite accurate apart from this bias.

**Recognition Accuracy ANOVA:** A 4x2x2 analyses of variance were performed on the recognition accuracy data from all four experiments. The judgment accuracy data yielded two significant main effects for experimental condition and sentence with \( F(15, 320)=48.0, p<0.005 \) and \( 59.6, p<0.005 \) respectively. The main effect for voice neared significance with \( F(15, 320)=3.32, p=0.07 \). Along with the extensive t-tests discussed above, the results of the ANOVA indicate that the different study tasks that were used in each experimental condition did produce changes in recognition accuracy. Additionally, the previous t-tests also preempted the main effect for sentence as the difference in accuracy for old and new sentences was significant in all the previous analyses. The lack of any voice effects obtained for new sentences within 3 of the 4 experiments explains the near-significance of the voice main effect even though old sentences consistently showed an effect of voice. The 3-way interaction was not significant in this analysis.

**Between Experiment t-Tests:** Figure 7 shows the average d’ values for each condition. Figure 8 shows the average d’ values compared across experiments.
Figure 7: Old and new sentence discriminability (d') grouped by Experiment. Probability boxes span the columns whose difference is significant with that probability from paired t-tests.

Figure 8: Old and new sentence discriminability (d') grouped by talker voice. Probability boxes span the columns whose difference is significant with that probability from unpaired t-tests.
In comparison to the average d’ scores from Experiment 1, both average d’s were significantly smaller for Experiment 2. The average d’ scores in Experiment 1 were 1.08 and 1.01 larger than the average d’s in Experiment 2 for old voiced and new voiced sentences respectively. An unpaired t-test between old voiced sentence d’ scores in Experiment 1 with old-voice d’s in Experiment 2 was significant, \( t(40)=4.89, p<0.001 \). The difference between new voiced d’s in the two experiments was also significant, \( t(40)=3.76, p<0.001 \). These results suggest that subjects were significantly less sensitive to sentence change across voice condition in Experiment 2 than Experiment 1. This pattern supports the hypothesis that the transcription task encouraged elaboration and a high degree of conceptual processing, which both aided later recognition.

Both of the average d’ scores from Experiment 3 were significantly lower than average d’s and 2 in unpaired t-tests: \( t(40)=3.97, p<0.005 \) for old voice d’ and \( t(40)=4.26, p<0.005 \) for new voice d’. This indicates that the word-pair transcription task had the effect of lowering sensitivity to sentence change. A likely explanation for this phenomena is that subjects were less able to remember whether the content words of the sentence they were judging appeared in a sentences during the study phase (and hence the sentence was old) or whether the content words had appeared in the word pair list (and hence the sentence was not necessarily old).

Like the accuracy scores, the average d’ scores for Experiment 3 were not significantly different from Experiment 3’s average d’ scores. These scores were both significantly lower than Experiment 2’s average d’ scores in unpaired t-tests: \( t(40)=4.1, p<.005 \) for old voice d’ and \( t(40)=3.9, p<.005 \) for new voice d’. Like the word-pair transcription task, the voice-monitoring task also had the effect of lowering sensitivity to sentence change. It is likely that subjects may have paid only enough attention to the study-phase sentences to judge gender. With more of their attention on voice quality and less on the lexical and semantic content of the sentences during study, subjects may have had a harder time at study recognizing both old and new sentences.

d’ ANOVA: A 4x2 ANOVA performed on d’ scores from all four experiments yielded two significant main effects for experimental condition and sentence type with \( F(7, 160)=91.5, p<0.005 \) and \( 8.8, p<0.005 \) respectively. As in earlier t-tests, the ANOVA of d’s indicates that the study-phase task had a significant effect on how well subjects could distinguish old sentences from new sentences. The second main effect likewise confirms the earlier t-test results that found significant or differences between old and new sentence d’s in all four experiments. The 2-way interaction was not significant.

General Discussion

Across four recognition memory experiments using several different study tasks, we observed consistent voice effects for sentences spoken in old voices. Subjects more accurately recognized “old” test phrase sentences when these sentences were presented at test in the same voice as their study presentation. This effect was represented in the accuracy data by a difference of 6.2% to 13.7% between OO sentences and ON sentences in the four conditions. These effects are particularly striking because subjects were never told explicitly that a memory test would follow their initial study of stimuli nor were they told to attend to the talker’s voice in any conscious, deliberate way - except in Experiment 4 where subjects were required to categorize the talker’s gender as male or female. Thus, we can conclude that rather than being filtered from the speech signal during processing, voice information was encoded and retained in the memory for sentences automatically without the conscious intent of the listener.

New sentences failed to display effects for voice in all except the final experiment that explicitly called subjects’ attention to a major component of voice information. As mentioned earlier, this pattern of results can be explained by considering the contribution of novel lexical and semantic content of the new
sentences. Few main nouns or verbs appeared more than once in any of the 80 stimulus sentences. Thus, with five content words in each sentence, subjects need only have noted the novelty of one or two of these words in order to recognize that the sentence was novel. Add to this the novelty of the overall meaning of the sentence that was different for each of the 80 stimulus sentences and subjects were unlikely to mistake a new sentence for an old one no matter what the voice. This may explain the overall bias to respond “new” across all four experiments. All conditions showed a greater accuracy for NN sentences than NO sentences even though this difference was not statistically reliable. These findings additionally support the proposal that listeners encode voice information in speech and that this information can be utilized to aid performance in a recognition memory experiment.

Perhaps the most interesting overall result of this series of experiments is the trend for recognition accuracy and discriminability to decrease while the effect of voice increased. This effect was evident in both new and old sentence judgments although old sentence judgments showed the phenomenon more clearly. From Experiment 1 to 4, judgment accuracy decreased from an OAA of 87.5% to an OAA of 66.4%. Average d’ likewise decreased from an average old voice d’ of 3.56 in Experiment 1 to an average of 1.54 in Experiment 4 – a drop of more than half across these experiments. While both these scores decreased, the difference in accuracy between old-voiced and new-voiced old sentences increased from 6.2% to 13.7% - a two-fold increase. Thus, while the changes in the study-phase task reduced the ability of subjects to explicitly recognize sentences during test, these same procedures served to increase the implicit effects of voice on sentence discriminability.

Experiment 1 showed accuracy near ceiling levels and can thus be considered the upper limit for subject’s ability to recognize the study phase sentences. When subjects were no longer required to transcribe the sentences – and hence were not encouraged to focus on the lexical and semantic content of the sentence – their overall recognition accuracy decreased. Since subjects who did not transcribe sentences would be less able to distinguish old from new sentences as subjects who could rely on lexical/semantic content, they may have relied on voice information to support recognition. This account explains the pattern of results obtained in Experiments 1, 2 and 3 well because voice information would likely play a more significant role in recognition when lexical information was reduced or attenuated. Experiment 4 may have encouraged the least amount of lexical information encoding since subjects needed only to recover talker gender from the stimuli.

When explicit attention is directed to talker gender and hence to correlated voice information in the speech signal, subjects may have encoded a comparatively rich representation of the talker’s voice information as compared to encoding of lexical information. This trend in the results suggests a flexible cognitive recognition system that can differentially utilize a variety of information when making judgments depending on the memory data available. The flexibility of this system is also characterized by the potential to process and encode a variety of information about speech including lexical/semantic and voice information.

The present discussion can best be concluded with consideration of the shortcomings of the present experiment and some directions that future experiments can take. First, although the stimulus file creation employed several layers of randomization when picking stimulus set files, the relative distinctiveness/similarity among the talkers and sentences was never formally measured. Thus, we cannot be certain that some stimulus set files were not composed of more or less similar talker and sentence combinations than others. Though a lexical analysis of the sentences for this purpose seems rather extraneous, an analysis of the perceptual similarity of talkers would be worthwhile. An important method for assessing this similarity would be a multi-dimensional scaling (MDS) analysis of talkers from this database. A pooling of subject’s similarity judgments for speaking sample stimuli in a forced choice ABX task could generate the data necessary for the MDS analysis. If successful, the MDS analysis could
provide a measure of perceptual similarity between talkers and hence allow us to control this factor in stimulus set creation for future experiments.

A second shortcoming of these experiments resides in the ambiguity in the instructions of what characterized an old sentence for the purposes of recognition. Subjects were simply told that an old sentence was one they heard in the study-phase of the experiment. No subject asked for any further clarification, but the fact remains that different subjects may have taken “heard at study” to mean both the voice and the sentence heard at study. We can be reasonably certain that no subject consistently judged sentences based on an old voice and old sentence criterion since OO and ON sentences differed in recognition accuracy by less than 15% in all conditions.

We cannot, however, rule out the possibility that some subjects inconsistently used an old voice/old sentence criterion during their recognition responses. If subjects made some recognition judgments based only on sentence type and some judgments based on both sentence and voice, then we could expect the small but significant differences between recognition accuracy for OO and ON sentences. We neglected to disambiguate the meaning of old sentences precisely because we could not design a way to communicate this distinction to subjects without calling attention to the fact that in all of the experiments the voices changed between study and test. Future experiments could benefit greatly from modifying these instructions that make the meaning of old sentence clear without calling attention to voice information. An additional safeguard against the ambiguity of the “old” criterion could be the inclusion of a post-experiment questionnaire. Such a questionnaire could probe each subject's understanding of “old” and thus their judgment criterion. Subjects who used an incorrect criterion, such as one that required old voice to be considered “old,” could be removed from the analysis. Our continued research in this area will incorporate such a post-experiment questionnaire.

With regard to the stimuli themselves, a richer corpus of sentences, particularly one that contained multiple tokens of the same talker speaking the same sentence, could add to the reliability of the pattern of results and conclusions discussed here. Since the OO sentence stimuli were actually the same stimulus files used in the study phase, one could question whether voice information was being used at all. Subjects could simply have based their recognition on some measure of overall acoustic similarity that would have been high for multiple presentations of the same stimulus file. Many subjects did, however, report that they were aware that some sentences were spoken by different speakers in the test phase. Clearly, at least for these subjects, their memory for sentences from the study phase included specific information about the voice of the talker.

The present series of experiments could also have benefited from additional data such as confidence ratings about the sentence recognition judgments. If subjects were told to rate their confidence that a sentence was old or new as well as the recognition judgment itself, additional converging measures could have been derived from the original recognition response. Additional data could also have been gathered after the test-phase such as subjects’ estimates of the number of talkers in the stimulus sets. Such measures could shed light on the individual differences in subject performance and could provide additional data about the accuracy of these estimates and how they may be systematically related to recognition performance across conditions.

The initial study-phase instructions could be altered to create an explicit rather than implicit recognition memory experiment. If subjects were told that a memory test would follow, they may adopt quite different encoding strategies during the study phase that could affect voice information utilization at test. If subjects explicitly rehearsed each sentence during the study phase to improve retention, then surface features such as voice information would likely degrade with successive rehearsals. A series of
experiments in which subjects are given differential amounts of time between stimuli to rehearse could uncover whether this strategy might be used.

Finally, the explicit nature of the present memory experiment could be changed to an implicit memory task such as perceptual identification that does not require conscious explicit recollection. Schacter (1987) has suggested that implicit memory is particularly sensitive to stimulus variability and to the perceptual encoding process. If surface perceptual features do exert more of an effect in an implicit memory task, then we would expect even larger effects of voice than those observed in the present set of experiments. The effects of surface information in sentences are even less explored than the small collection of research on explicit sentence memory. Thus, implicit memory experiments represent an important avenue for future research on the encoding of episodic or instance specific properties of sentences and the talkers that produce them.

While the problems mentioned earlier do call into question any specific conclusions about voice effects and recognition accuracy, the general conclusion that spoken language processing retains talker-specific information for encoding stands. This encoding of extra-linguistic information takes place without the conscious intent of the listener. To the extent that talker voice was never mentioned during any of the four experiments yet subjects still showed consistent voice effects, we find evidence in support of incidental encoding of talker-specific, extra-linguistic information. Additionally, since neither voice information nor a test of memory for the sentences was ever mentioned to subjects, it is unlikely this incidental encoding of voice information is in some way an artifact of the experimental procedure. Thus, it is probably the case that listeners automatically and incidentally encode talker-specific information during “everyday” speech processing.

The retention of rich variation in the memory representation of the speech signal means that perceptual processing does not strip away variation in the speech signal before encoding into memory. Thus, our memories of speech events are not idealized sequences of words or phonemes that preserve the gist of what was said but not the specific variation present in the stimulus. Rather, our memories for speech events preserve at least some of the specific, rich variation present at original encoding. If perceptual processing does not filter variation from the signal before encoding into memory, then we can fairly question whether “variation filtering” of the speech stimulus occurs at all. If we abandon the notion that speech recognition requires, at some stage, an idealized, formal representation of the speech stimulus to begin with, then we need not explain how variation can be preserved for encoding on the one hand, but filtered away during the recognition process on the other. That is, the regularities in the multi-modal information for speech may not be interpreted in the form of idealized units extracted from the variable speech signal.

The process of speech recognition could extract higher-level regularities from the speech signal that depend upon analysis of variation in the signal caused by the special circumstances and characteristics of the talker and listener. While no current theory models speech recognition outside a framework incorporating a formalized, idealized phonemic or segmental stage, it should no longer be taken for granted that speech recognition is a normalizing, abstracting process. The variability in the speech signal may be as important to our understanding and encoding of speech as the regularity of the phonemes and words that have traditionally been considered the nucleus of speech recognition.
References


Summerfield, Q. & Haggard, M. P. Vocal tract normalization as demonstrated by reaction times. *Report of Speech Research in Progress, 2(2)* Queens University of Belfast (pp. 12-23).