

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

The Hoosier Audiovisual Multi-Talker Database¹

Sonya Sheffert, Lorin Lachs and Luis R. Hernández

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

¹ This research was supported by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University Bloomington.

The Hoosier Audiovisual Multi-Talker Database

Abstract. This report describes the Hoosier Audiovisual Multi-Talker Database, a 3000 word multimodal video database developed at the Indiana University Speech Research Laboratory. The corpus consists of ten adult talkers (five male and five female) producing 300 familiar monosyllabic English words. Each spoken word is presented as a dynamic full-motion color movie. The database also includes information about the lexical characteristics and the intelligibility of each word.

Objectives of the Database

The objectives of our research program on multimodal perception are to: 1) investigate the perception and integration of auditory and visual information during multimodal language processing, 2) explore the nature of the memory traces for spoken words produced by different speakers, 3) assess the effects of optical and auditory talker information on speaker normalization, 4) examine the relationship between implicit and explicit memory for voices, faces, and spoken words, 5) investigate the effects of talker-specific characteristics on multimodal speech intelligibility, and 6) determine whether structural characteristics of the mental lexicon contribute to listener's ability to use multimodal information during the course of word recognition. In order to achieve these goals, it was necessary to create a large digital database of spoken words.

Organization of the Database

The organization of the database is based on lexical neighborhoods, or collections of words characterized by acoustic-phonetic similarity (Luce, 1986; Luce & Pisoni, 1998). There are 150 words designated as "easy", and 150 as "hard". Easy words reside in sparse neighborhoods populated by only low frequency words, whereas hard items reside in dense neighborhoods shared by many high frequency words.

The database is organized around acoustic-phonetic similarity because previous research has demonstrated that these two dimensions, frequency and similarity play an important role in spoken word identification. Target items from low-density, low-frequency neighborhoods are identified faster and more accurately than hard words, presumably because there is less lexical competition among the items. Differences in the identifiability of easy and hard words have been demonstrated in normal listeners (Luce, 1986; Luce & Pisoni, 1998), in hearing impaired listeners and patients with cochlear implants (Sommers, Iler-Kirk, Pisoni & Osberger, 1993), young children (Charles-Luce & Luce, 1990) and older adults (Sommers, 1996). The development of Hoosier Audiovisual Multi-Talker Database allows us to extend the findings by addressing the effects of lexical confusability on multimodal language processing.

Because each word was produced by ten different speakers (five male and five female), the database can also be organized by talker. The use of several talkers permits us to build on a growing literature demonstrating that speech perception and spoken word recognition are influenced by stimulus variability and talker familiarity (for a review, see Pisoni, 1993).

The tokens in the databases are dynamic full-motion color movies rather than as static photographs. This is important for exploring issues concerned with the integration of multimodal information; that is, rather than having audio information which is tied to static visual information arbitrarily, the database contains two tracks of information from different sensory modalities which are

lawfully tied to one another. This will allow users of the database to investigate questions concerning the multimodal encoding and processing of speech in a more ecologically valid manner.

Creating the Hoosier Audiovisual Multi-Talker Database

Methods

Subjects

Ten different talkers (five males and five females) were recruited from the Indiana University community. The age of the talkers ranged from 18 years to 32 years, and they represented a range of geographical areas in the Midwest: Chicago (N=4), Indiana (N=1), Iowa (N=1), Oklahoma (N=2), St. Louis (N=1) and Wisconsin (N=1). None of the talkers wore glasses. One talker had a small mustache and beard at the time of the recordings. All talkers were Caucasian.

Stimulus Materials

Each talker was videotaped while producing 300 monosyllabic CVC words. The words were selected from the Hoosier Mental Lexicon (Nusbaum, Pisoni, & Davis, 1984), which is based on a 20,000 word on-line dictionary. The HML database provides information on word frequency (Kucera & Francis, 1967), word familiarity (Nusbaum et al., 1984), neighborhood density and neighborhood frequency. Together, these variables allowed one to specify the relative degree of confusability among the items, and to separate the words into "easy" and "hard" items.

Only highly familiar words (6 or greater on a 7-point scale) were selected for the Hoosier Audiovisual Multi-Talker Database. Neighborhood density or the number of phonetically similar words or neighbors of a target item, was determined by a one-phoneme substitution, addition, and deletion metric (Greenberg & Jenkins, 1964). Mean neighborhood density was 18 for the easy items and 24 for the hard items. Neighborhood frequency, or the average frequency of all the items within a target neighborhood, was obtained from Kucera and Francis (1967). Mean neighborhood frequency was 41 per million for the easy items and 251 per million for the hard items. A summary of the lexical characteristics of the items is displayed in Table 1.

Table 1.

Lexical characteristics of the easy and hard words.

	Easy	Hard
Mean Lexical Frequency	319	28
Mean Familiarity	7	7
Mean Lexical Density	18	24
Mean Neighborhood Frequency	41	251

Talkers were videotaped in a sound-attenuated professional recording studio using an 8mm professional SVHS Canon video camera. Each word was presented in isolation on a CRT screen at a fixed citation rate of 1 word every 3 seconds. Each talker received a different random ordering of the words.

During the videotaping, all the speakers wore a solid black shirt. Talkers were instructed to speak clearly and naturally at a normal conversational rate. They were told to look directly into the camera while assuming a neutral facial expression and to avoid any extraneous head to body movement. They were also instructed to begin and end each utterance with their mouth closed. Words which were mispronounced or accompanied by extraneous movement were re-recorded.

The video images were captured, digitized and segmented using a commercial software package (Adobe Premiere) installed on a Macintosh Quadra 950. The video equipment was designed to record and playback full-motion video at 30 frames per second (NTSC standards). The audio signal was digitally sampled at 22kHz with 16-bit resolution. The video was digitized at 30 fps, with 24-bit resolution at 640 by 480 pixel size. Each word was made into a free standing Quicktime movie using Adobe Premier's movie making function. The movies were created in Radius format and converted to TARGA format in order to maintain compatibility with the latest video hardware. The audio signal for all tokens across talkers was equated for root mean square amplitude (RMS). All movies were leveled at 54 dB. The resulting leveled movies were stored on digital optical disks.

Each token is approximately 2 seconds long. Every spoken word is buttressed by approximately .5 seconds of silence during which the talker's mouth was closed. The overall integrity of each movie was assessed by two trained listeners (a phonetician and a psychologist). Each token was screened for the presence on any of the following errors: Acoustic distortion, background noise, ambiguous or incorrect pronunciation, nonspeech mouth sounds; unusual lip movement, unusual eye movement, and visual distortion. The final version of the movies are presented as a full-screen image (640 x 480) in 24-bit color and presented on a high resolution 17" monitor.

Speech Intelligibility

Future users of the database may need data on the intelligibility of each of the tokens. Therefore, each stimulus was tested for ease of perceptual identification. Results showed that the majority of tokens in the database were highly intelligible with little variability between talkers. We report only the AV intelligibility in this report. Assessment of the intelligibility of these stimuli in the visual- and audio-only contexts is currently in progress.

Subjects

100 Indiana University undergraduates participated in return for partial course credit in an introductory psychology course or for five dollars. All subjects had normal hearing, were native English speakers, and reported no history of speech or hearing disorders at the time of testing.

Materials

One Macintosh Quadra 950, one Macintosh PowerPC 7100, and three PowerComputing 180s, each with a 17" Sony Trinitron Monitor, were used to present the video displays of the stimuli. The stimuli consisted of the movies from all ten talkers.

Procedure

A computer program was used to control stimulus presentation and collect subject response. Each subject was presented with a randomly ordered set of movies; each set of movies consisted of all the tokens

spoken by a particular talker. Stimuli were presented using 17" monitors and BeyerDynamics DT-100 stereophonic headphones at a loudness of 75 dB/SPL. Before the presentation of the first stimulus, subjects were given a set of typed instructions explaining the task and procedures. Listeners were informed that they would see a series of movies in which a person would speak a single English word. Subjects were informed that they would be hearing each word only once. After each stimulus, subjects were required to identify the word by typing its spelling on the keyboard. Subjects were instructed that the next movie would not be presented until they pressed the RETURN key. They were also reminded to take time to make sure that the response they typed was the response which they intended to make before entering it. Each response was then collected in a text file which contained the name of the movie, it's order in the presentation, and the subject's response. Each subject, therefore, had his/her own logfile.

Data Analysis

All logfiles from subjects who viewed a particular talker were analyzed in tandem. The intelligibility data program scanned the responses made across subjects for each movie. Responses that matched the intended response were scored as correct. Any strings which were homophonic with the intended response or any strings which matched the correct response except for obvious typos were also accepted as correct. All other responses were marked as incorrect identifications of the target word. The data were analyzed for the effects of three factors on intelligibility: talker, easy/hard lexical classification, and word.

Results

The overall intelligibility of the words, expressed as the percent correct identification of all words from all talkers, was 98.57%. The mean intelligibility of each word across talkers was also calculated and a table of these scores was constructed for further reference. This table will allow users of the database to have readily available statistics on the intelligibility of the words contained therein and will allow the use of relative intelligibility as a possible experimental variable in the future. Closer examination of the table reveals several relevant characteristics of the database as a whole. The lowest average intelligibility was 74.09% (for the word "cot"). However, most items produced intelligibility scores that were near ceiling. In fact, examination of the distribution of scores reveals that 99.97% of the words were identified across talkers with greater than 90% accuracy.

Figure 1 shows the intelligibility scores for each talker separated by lexical category. A 2x10 ANOVA (Lexical Category, Talker) performed on the items revealed a main effect of Lexical Category [$F(1,3010) = 5.4$; $p=0.02$]. On average, Easy Words were identified 0.5% better than Hard words. However, there was no main effect of talker [$F(9,3010) = 0.3051$; N.S.], and no significant interaction between these two variables [$F(9,3010) = 0.525$; N.S.]. Post-hoc Fisher's PLSD pairwise analysis of the talker combinations showed no significant difference between any two specific talkers. In addition, a one-way ANOVA by Sex of the Talker showed no significant effect of this variable [$F(1,9) = 0.112$; N.S.]. These results are not surprising in light of the fact, noted above, that most of the intelligibility scores were near ceiling levels.

 Insert Figure 1 About Here

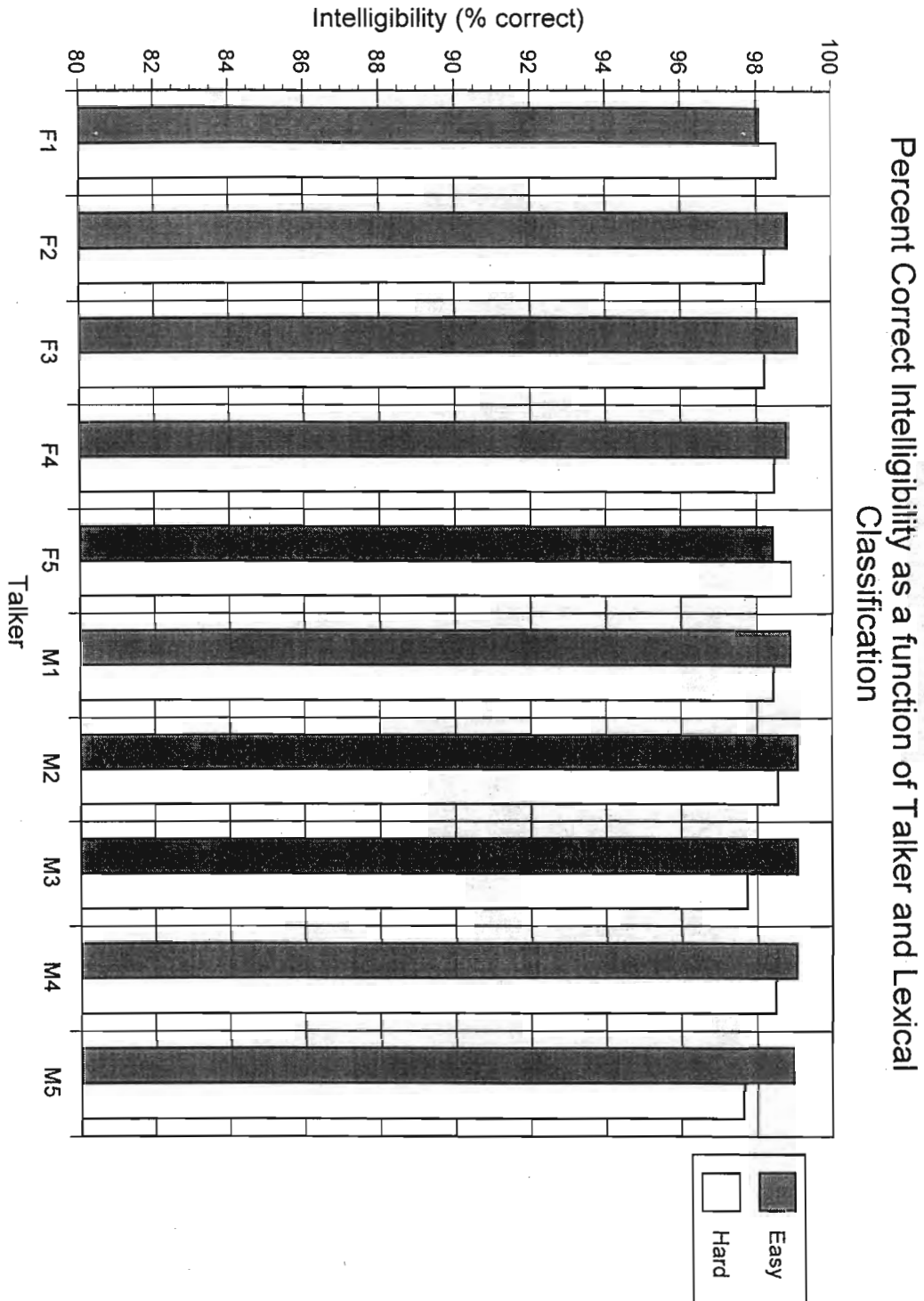


Figure 1. Mean percent correct intelligibility of "Easy" and "Hard" words, displayed as a function of speaker. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

Conclusion

The development of the Hoosier Audiovisual Multi-Talker Database was completed in two phases. In the first phase, the stimuli were filmed, digitized, leveled and stored. During the second phase, the stimuli were subjected to a test of intelligibility. All in all, we was found that there were no significant differences between stimuli across the talker who spoke them. In addition, there were no significant differences in intelligibility as a function of the words contained in the stimuli. There was, however, a significant difference between stimuli classified as Easy or Hard lexically. Although it may at first glance appear detrimental that differences between stimuli were not found to be significant, it should be remembered that the vast majority of the stimuli contained within the database had intelligibility scores near ceiling. Overall, more than 99% of the tokens were identified with greater than 90% accuracy. This is most likely the reason for a lack of significant effects, but it at the same time attests to the fact that the construction of the Hoosier Audiovisual Multi-Talker Database has produced a large corpus of stimuli which are highly intelligible and are readily available for use in future investigations using audiovisual stimuli.

References

- Charles-Luce, J. & Luce, P. A. (1990). Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language*, *17*, 205-215.
- Greenberg, J. H. & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, *20*, 157-177.
- Kucera, F. & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Dissertation Abstracts International*, *47* (12-B, Pt. 1), 5078.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, *19*, 1 - 36.
- Nusbaum, H. C., Pisoni, D. B. & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report No. 10* (pp. 357-377). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, *13*, 109-125.
- Sommers, M. S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken word recognition. *Psychology and Aging*, *11*, 333-341.
- Sommers, M. S., Kirk, K. I., Pisoni, D. P & Osberger, M. J. (1993). Some new directions in evaluating the speech perception performance of cochlear implant patients. A first report. In *Research on Spoken Language Processing Progress Report No. 19* (pp. 271-281). Bloomington, IN: Speech Research Laboratory, Indiana University.