

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 22 (1998)
Indiana University

Update: The Hoosier Audiovisual Multi-Talker Database¹

Lorin Lachs and Luis R. Hernandez

*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. Special thanks go to Andrew Bangert for his assistance during data collection and compilation.

Update: The Hoosier Audiovisual Multi-Talker Database

Abstract. This report describes continued work on the intelligibility of the stimuli contained in the Hoosier Audiovisual Multi-Talker Database (Sheffert, Lachs, & Hernandez, 1997). All the stimuli in the database have now been tested using three different presentation conditions: audio-visual, audio-only, and visual only. A re-analysis of the audio-visual data, together with new analyses of the data obtained in the unimodal conditions, revealed significant effects of presentation modality. The data are presented as summary statistics for use in the selection of stimuli from the database.

Introduction

A previous paper (Sheffert, Lachs, & Hernandez, 1997) described the construction of a 3000 token stimulus database consisting of digitized movie clips, each of which contained one of ten talkers speaking one of 300 words. Our preliminary report summarized the findings from an intelligibility study on the database, in which each token was presented to 10 listeners for identification.

Sheffert et al. (1997) reported that the overall intelligibility of the words was 98.57%, with a range of 74.09% to 100%. In addition, a per-item analysis showed that most of the tokens were identified near ceiling, with 99.97% of the tokens in the database being identified across talkers with above 90% accuracy. Finally, Sheffert et al. (1997) report a main effect of lexical category. Lexically “easy” words were identified more easily than lexically “hard” words (Luce, 1986; Luce & Pisoni, 1998). No differences were found between overall intelligibilities across talkers.

In all previously reported cases, identification was made using audio-visual presentation of the clips. That is, the findings in Sheffert et al. (1997) concern identification by subjects when they had access to both the audio and video signal for each token. The present study reports further tests of the intelligibility of the database using audio-only and visual-only presentation.

Method

Subjects

Three hundred and six Indiana University undergraduates participated in exchange for 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. One hundred subjects participated in the audio-visual identifications (Sheffert et al., 1997), 102 of the subjects participated in audio-only identifications and 104 subjects participated in the visual-only identifications.

Materials

Two Apple Macintosh and three Macintosh clone computers each equipped with a 17” Sony Trinitron Monitor (0.26 dot pitch) and its own video processing board were used to present the stimuli to subjects. The video processing boards were each capable of handling clips digitized at 30fps with a size of 640 x 480 and 24-bit resolution. The stimuli consisted of all the movies from all the talkers in the database.

Procedure

As in Sheffert et al. (1997), a computer program was used to control stimulus presentation and collect subject responses. The custom-designed software for presentation of the digitized movies was altered such that the video track for each movie was not displayed in the audio-only condition; alternatively, the software was altered such that the audio track for each movie was not displayed in the video-only condition. Thus, the same stimuli and presentation program were used across the Sheffert et al. (1997) study and the two additional conditions in the present report with minor alterations to the program being necessary for unimodal presentation.

Each subject was presented with a randomly ordered list of movies for identification with each list of movies consisting of all the tokens spoken by a particular talker. Stimuli in the audio-only condition were presented over BeyerDynamics DT-100 stereophonic headphones at a loudness of 75 dB/SPL. Stimuli in the video-only condition were presented over the Sony Trinitron 17" monitors.

Before the presentation of the first stimulus, subjects in both conditions were handed a set of typed instructions that explained the task and procedures. Listeners were informed that they would hear a series of stimuli or see a series of video clips in which a person would speak a single English word. Subjects were informed that each stimulus would be presented 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 stimulus 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 that they intended to make before entering it. Each response was then collected in a logfile that contained the name of the movie, its order in the presentation, and the subject's response.

Data Analysis

All logfiles from subjects who viewed a particular talker were analyzed in tandem. An intelligibility data program scanned the responses made across subjects for each movie. Responses that matched the intended response were scored as correct. Any strings that were homophones of the intended response or any strings that 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. Talkers were given labels according to their gender. Thus, M1 was the label given to the first male talker, while F3 was the label given to the third female talker. The data were analyzed for the effects of four factors on intelligibility: Talker (M1, M2, M3, M4, M5, F1, F2, F3, F4, F5), Lexical Category (Easy, Hard), Word, and Presentation Modality (audio-visual, audio-only, video-only). For convenience, the audio-visual condition will herein be referred to as the AV condition, the audio-only condition will be referred to as the AO condition, and the video-only condition will be referred to as the VO condition.

Upon examination of the AO and VO data, we discovered that the orthographic transcriptions of the intended utterances for several of the movies in the database were ambiguous (i.e., "READ", "LIVE"). Across the entire set of words, 37 tokens out of the total 3000 had this property. This prompted a re-examination of the stimulus files themselves. The result of this re-examination revealed that some of the talkers spoke these ambiguously spelled words using one pronunciation, while still others spoke them using another. Since our lexical classification of words is dependent on acoustic-phonetic similarity, not orthographic similarity, it turned out that some of the stimuli could not be categorized as either "easy" or "hard" words. That is, for those ambiguously spelled words that were uttered, during the initial filming of the stimuli, by a particular talker using the *non-intended* pronunciation (as determined by our pre-computed similarity data), a lexical classification could not be determined. These tokens will hereby be referred to as words in the "unknown" lexical classification. Although we have compiled intelligibility

data for these tokens, we have not entered these data into the final analysis, since we can not be certain as to their lexical status.

This new parsing of the stimulus database warranted a re-analysis of the identification data obtained in the audio-visual presentation condition (previously reported in Sheffert et al., 1997). As such, the intelligibility of tokens in all three presentation conditions is reported below. However, it should be noted here that the overall characteristics of the audio-visual condition reported previously were not (and could not) have been affected by this re-analysis. That is, the overall intelligibility, the range of intelligibility scores, and the proportion of the database whose intelligibility was above 90% remained constant across the reclassification of the ambiguous words.

Results

Audio-Visual Data

Figure 1 shows the AV intelligibility scores for each talker separated by lexical category (easy or hard). As can be seen from the graph, intelligibility scores for easy words were higher than intelligibility scores for hard words for all talkers except for F5. Indeed, a 2x10 ANOVA (Lexical Category, Talker) performed on the items revealed a main effect of Lexical Category [$F(1, 2977) = 24.435, p \leq 0.0009$] and a main effect of Talker [$F(9, 2977) = 3.051, p \leq 0.001$]. There was no significant interaction between the two factors [$F(9, 2977) = 1.045, n.s.$].

 Insert Figure 1 about here.

In light of the ceiling level performance, it is perhaps not surprising that there was no interaction between the two variables.

Audio-only data

An analysis of the data from the audio-only condition revealed an overall intelligibility of 96.93%. The average intelligibility scores for words in this condition ranged from 46.06% ("been") to 100% intelligibility. Token intelligibility spanned the entire range from 0% to 100%. 92.65% of the tokens were identified with greater than 90% accuracy.

Figure 2 shows the AO intelligibility scores for each talker separated by lexical category. Easy words yielded higher intelligibility scores than hard words for all talkers. A 2x10 ANOVA (Lexical Category, Talker) was performed on the items. As with the audio-visual data, main effects of both factors were obtained [Lexical Category: $F(1, 2996) = 29.797, p \leq 0.0009$; Talker: $F(9, 2996) = 3.546, p \leq 0.0009$]. In addition, there was a significant interaction between the two variables ($F(9, 2996) = 2.118, p = 0.025$). That is, the extent to which the lexical variables played a role in identification was dependent on the talker. Post-hoc Tukey's HSD analyses by talker revealed significant differences in the average intelligibility of M1 when compared with the intelligibilities of F1 ($p \leq 0.001$), F4 ($p \leq 0.0009$), F5 ($p = 0.013$), M2 ($p \leq 0.0009$), and M4 ($p = 0.005$).

 Insert Figure 2 about here.

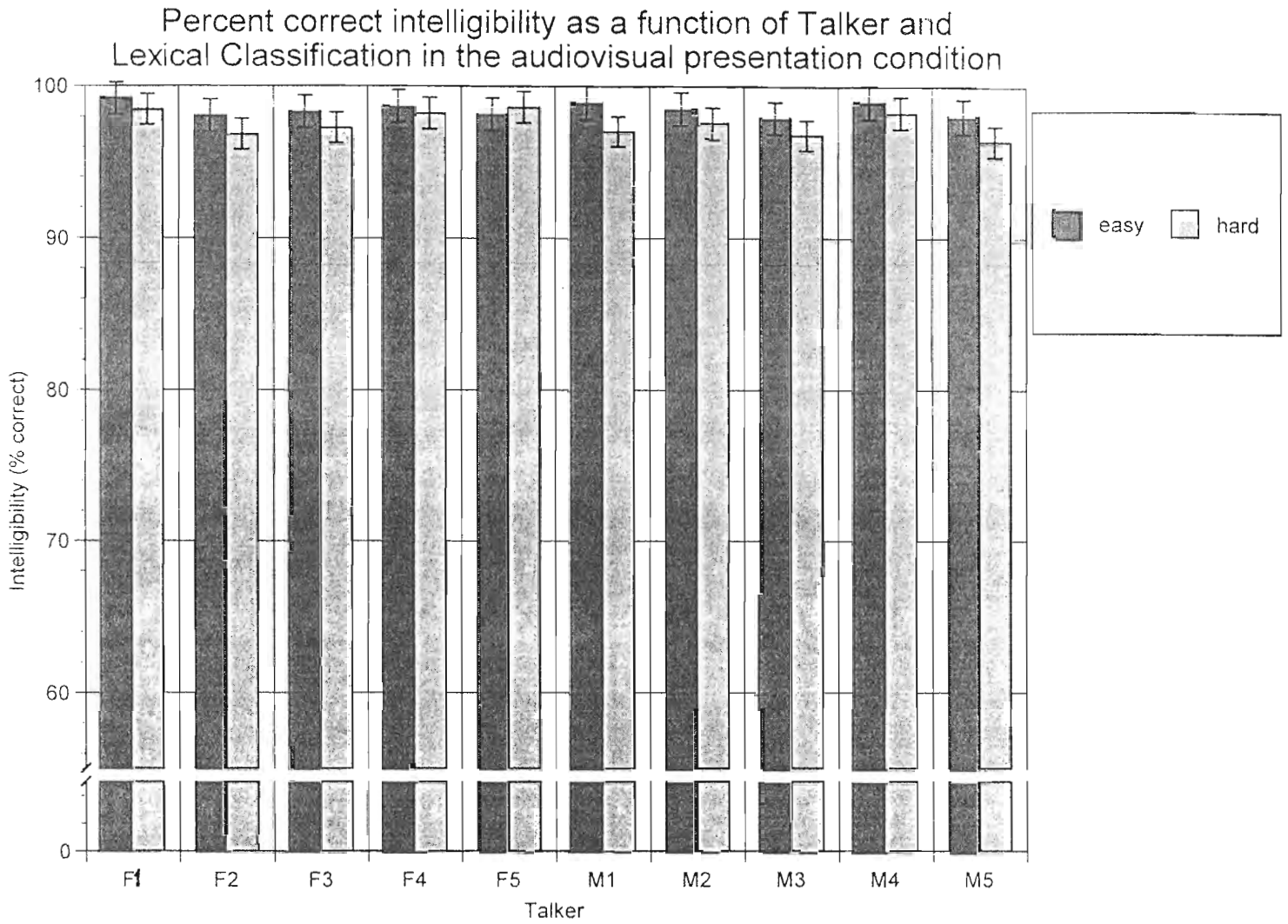


Figure 1: Mean percent correct intelligibility of “Easy”, and “Hard” words in the audio-visual condition, displayed as a function of speaker. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers. Error bars indicate standard errors.

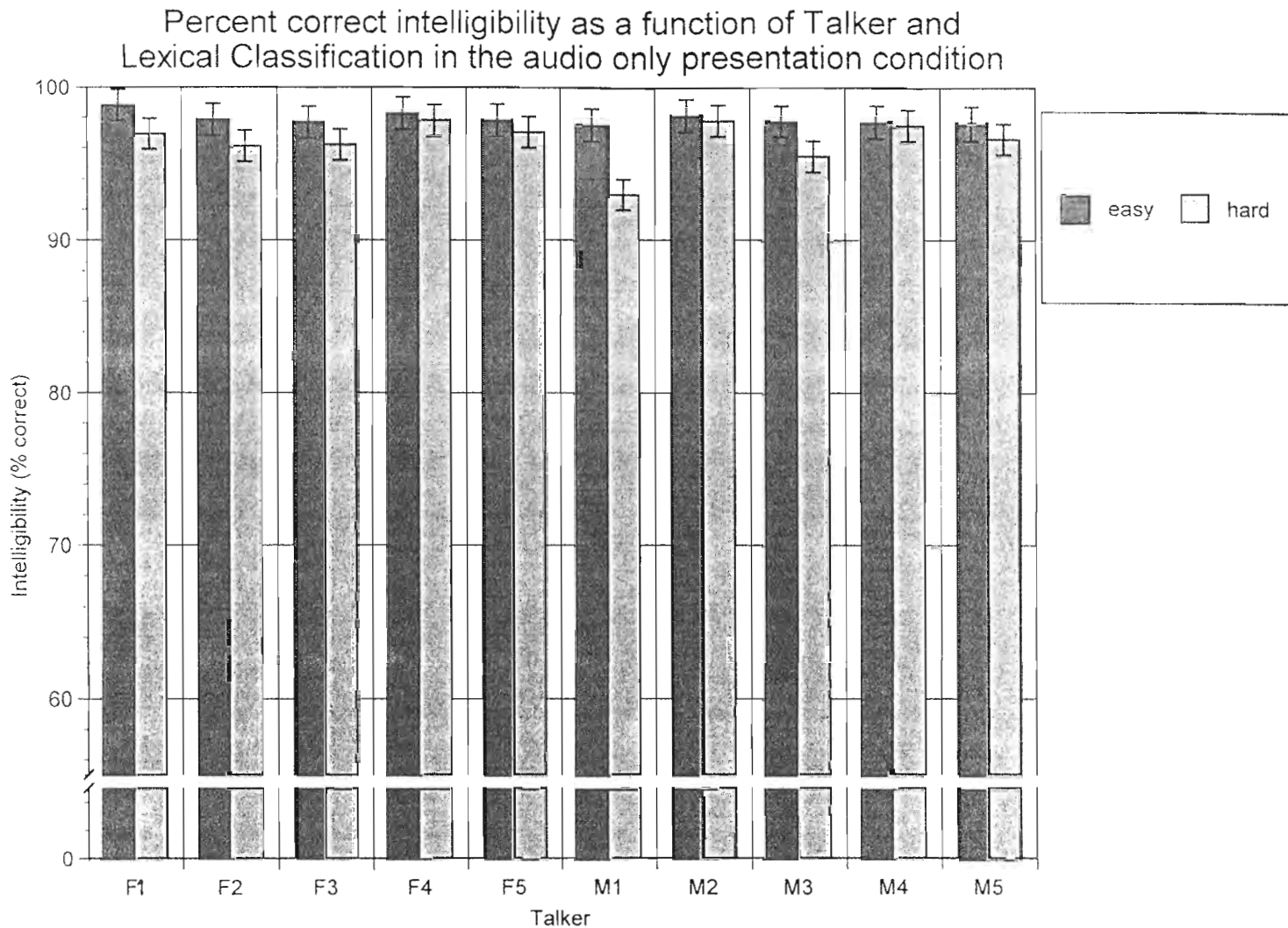


Figure 2: Mean percent correct intelligibility of “Easy”, and “Hard” words in the audio-only condition, displayed as a function of speaker. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers. Error bars indicate standard errors.

Video-Only Data

An analysis of the data from the video-only condition revealed an overall intelligibility of 14.13%. The average intelligibility scores for words in this condition ranged from 0 to 70.57% (“five”). Token intelligibility spanned the entire range from 0% to 100%. 0.9% of the tokens were identified with greater than 90% accuracy. 90% of the tokens were identified with less than 40% accuracy.

Figure 3 shows the VO intelligibility scores for each talker separated by lexical category. Once again, easy words were identified with greater accuracy than hard words, regardless of the talker. A 2x10 ANOVA (Lexical Category, Talker) was performed on the items. As with the other two presentation modalities, there was a significant effect of Lexical Category [F(1, 2943) = 279.131, p <=0.0009]. There was also a significant main effect of Talker [F(9, 2943) = 9.173, p <= 0.0009]. There was no significant interaction between the two variables (F(9, 2970) = 0.996, N.S.).

 Insert Figure 3 about here.

Table 1 presents the results of Post-hoc Tukey’s HSD analyses on the talker effect. An asterisk denotes that the comparison represented by the particular cell yielded a significant difference. As can be seen, there were differences in many of the pairwise comparisons

Table 1
Results of Tukey’s HSD post-hoc pairwise comparisons on
talker intelligibilities in the VO condition

	M1	M2	M3	M4	M5	F1	F2	F3	F4	F5
M1	0									
M2	*	0								
M3	*		0							
M4				0						
M5	*				0					
F1	*			*		0				
F2	*					*	0			
F3	*							0		
F4	*			*			*		0	
F5	*			*			*			0

All Three Presentation Modes

The data from all three experiments were submitted to a 3-way ANOVA (Presentation Mode, Lexical Classification, and Talker). The 3-way ANOVA revealed a significant main effect of presentation mode [F(2,8896) = 45239.89, p <= 0.0009], and a significant main effect of lexical classification [F(1, 8896) = 334.845, p <= 0.0009]. There was also a main effect of talker [F(9, 8896) = 12.507, p <= 0.0009].

Percent correct intelligibility as a function of Talker and Lexical Classification in the video only presentation condition

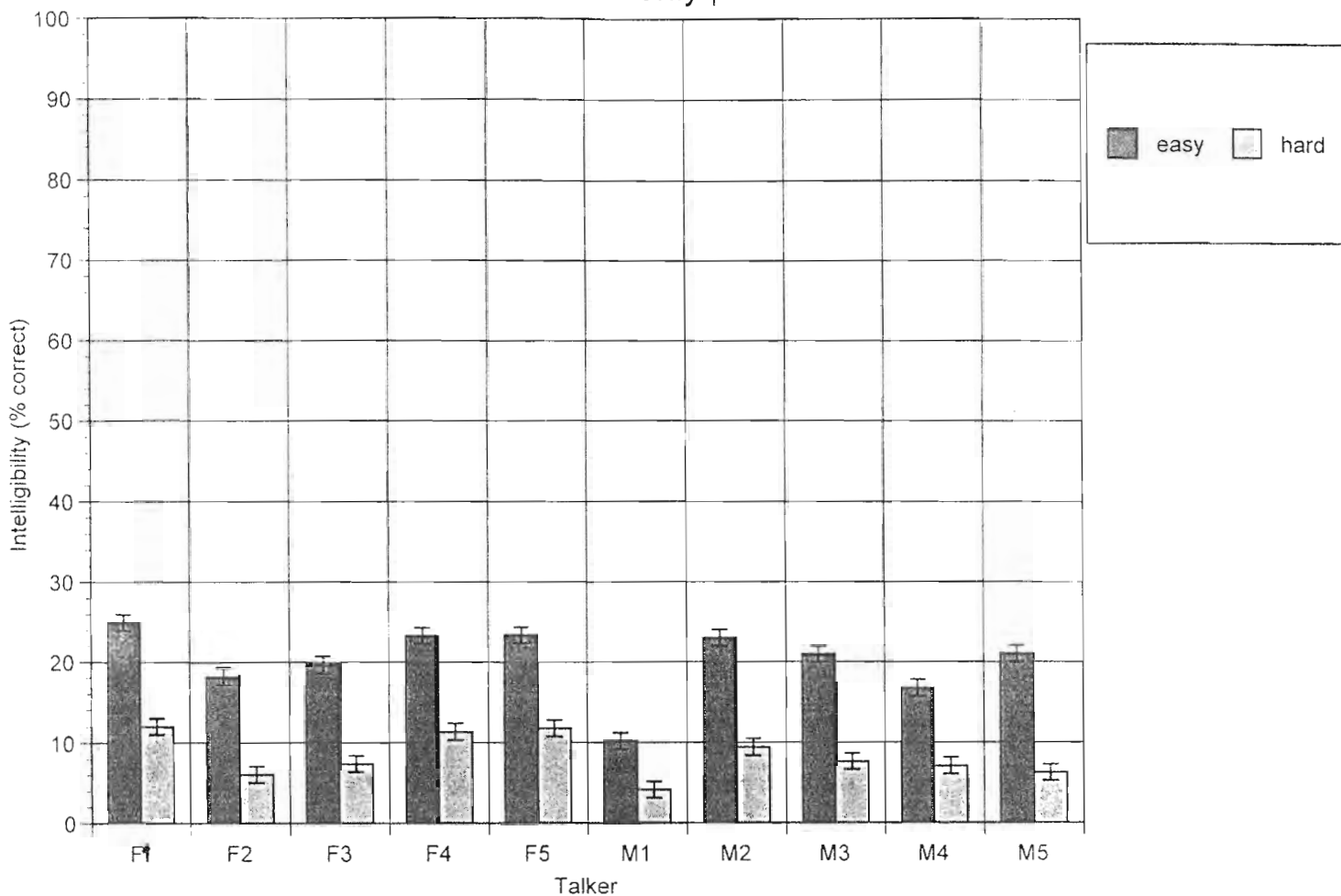


Figure 3: Mean percent correct intelligibility of “Easy”, and “Hard” words in the video-only condition, displayed as a function of speaker. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers. Error bars indicate standard errors.

Post-hoc Tukey's HSD tests on the main effect of presentation mode showed that audio-visual scores were significantly different from audio-only scores ($p = 0.01$), and that scores in the visual only condition were significantly different from scores in both of the other presentation modes at $p \leq 0.0009$. As expected, then, there were real differences in the intelligibilities of words presented across the three presentation modes.

All post-hoc comparisons between lexical categories were significantly different at $p \leq 0.0009$. Thus, across all three presentation modes, intelligibility scores in the lexical categories were distinct.

Presentation Mode also interacted with Lexical Classification [$F(2,8896) = 182.248$, $p \leq 0.0009$], indicating that a word's status as "easy" or "hard" differentially affected intelligibility judgements depending on the amount of information to which listeners had access. Figure 4 shows the average intelligibility of items, split by lexical classification, as a function of the presentation modality. From the graph it is clear that Lexical Category also had a differential effect across the video-only condition.

 Insert Figure 4 about here.

There was also a significant interaction between Presentation Mode and Talker [$F(18, 8980) = 5.952$, $p \leq 0.0009$]; that is, the intelligibility of the talkers themselves was affected by the presentation modality. Figure 5 shows the average intelligibility of items, split by talker, as a function of presentation modality. The graph clearly illustrates that the degree to which presentation mode affected intelligibility across talkers varied widely across the talkers. For example, the decrement in F1's intelligibility as a result of visual-only identification (relative to the intelligibility of F1 in the AV condition) is not as great as the decrement in M1's intelligibility due to visual-only identification (relative to M1's AV intelligibility).

 Insert Figure 5 about here.

There was no interaction between Lexical Classification and Talker [$F(18, 8980)=0.608$, N.S.].

Discussion

The results from the present intelligibility tests show that our goal of constructing a useful set of intelligible audio-visual stimuli has been met. A vast majority of the stimuli were highly intelligible under audio-visual and Audio-only presentation conditions. The stimuli also exhibit properties that correspond with existing data on the identification of words from lexical neighborhoods with different similarity properties (Luce, 1986; Luce & Pisoni, 1998). "Easy" words were, across all the conditions, identified more accurately than "hard" words.

Interestingly, the data from both the audio-visual and the audio-only conditions supported an interaction between the identity of the talker and the lexical category. The degree to which the properties of a particular talker interact with lexical properties like neighborhood density and frequency remains a largely unexplored question, and our data provide preliminary evidence in the investigation of these factors. Specifically, our results point to a role for talker-specific information in the process of lexical

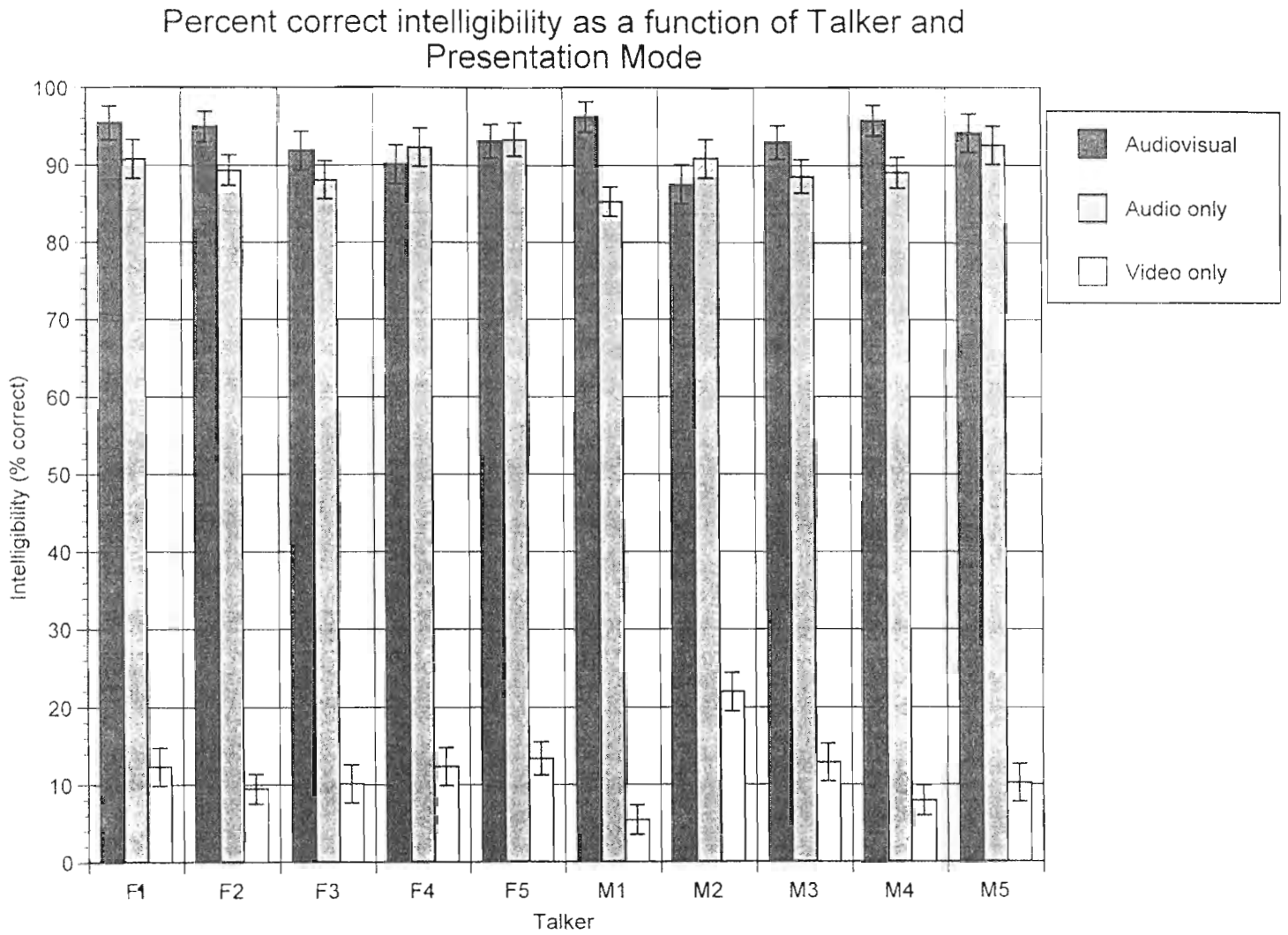


Figure 4: Marginal mean intelligibilities for the interaction between Lexical Classification and Presentation Mode. Error bars indicate standard errors.

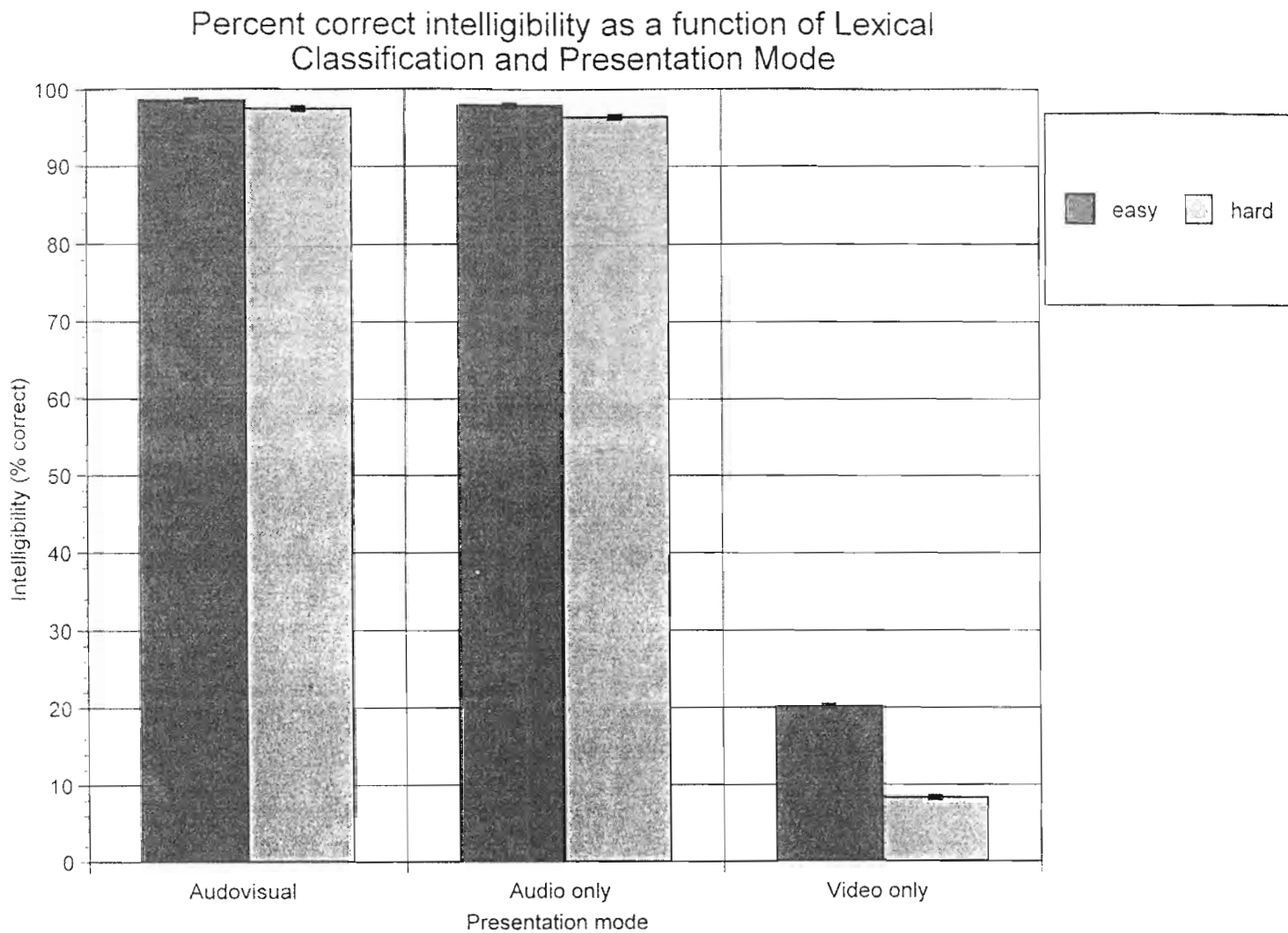


Figure 5: Marginal mean intelligibilities for the interaction between Talker Identity and Presentation Mode. Error bars indicate standard errors.

access. Given the growing evidence that lexical representations contain highly detailed information about vocal sources (see Goldinger, 1998 for a review), it is promising to see that these two traditionally disparate sources of variance can have interactive effects during lexical access. In addition, several interesting properties of these stimuli emerged when they were presented using visual-only information. Specifically, there was also an effect of Lexical Category on identification in the video-only presentation. This seems to imply that lexical characteristics rely not only on acoustic-phonetic similarity, but also upon similarity along dimensions that incorporate visual information. In addition, examination of Figure 5 reveals that Lexical Category has a larger effect on identification in the video-only condition than in the other conditions, although this is probably due to the ceiling effects in the other two conditions. We withhold speculation about the mechanism by which these effects manifest and further investigation will be forthcoming.

Finally, we note that a large spreadsheet of all the identification data from all three presentation modes has been compiled and is being stored in tandem with the stimulus files themselves. The spreadsheet contains summary statistics at all levels of analysis, from the average intelligibility of particular words spoken by particular talkers, to the intelligibility of words spoken across talkers, to the overall intelligibility of the stimuli across conditions. It is our hope that this spreadsheet will provide a useful way for picking stimuli relevant to particular experiments, and that it may also provide a source for further experimental analysis of the effects of presentation modality on intelligibility.

References

- Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105 (2), 251–279.
- 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.
- Sheffert, S., Lachs, L. & Hernandez, L.R. (1997). The Hoosier Audiovisual Multi-Talker Database. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 578-583). Bloomington, IN: Speech Research Laboratory, Indiana University.