INVARIANCE OF LEXICAL CONTEXT USE IN SPOKEN WORD RECOGNITION

Theresa Hnath-Chisolm
University of South Florida

ABSTRACT

Determining a sensory aid increases access to acoustic speech patterns typically involves measuring percent correct whole-word recognition for open-set monosyllabic word lists. Thus, not only is access to acoustic/phonetic information assessed, but also lexical knowledge. These effects can be separated by application of simple probability theory, in which the recognition probabilities for the wholes (P_w) is compared to those for the parts ($P_{p)}$. The result is a j-value indicating the number of independent perceptual units necessary for correct word recognition. An underlying assumption of ***j*** is that it is a constant - that it is not affected by the underlying recognition probabilities of the parts. This assumption was examined by varying type and availability of sensory information through: (1) stimuli presentation via three input-modalities; and, (2) presentation by multiple speakers. Results confirmed j-values were independent of speaker but not modality manipulations, suggesting the speech recognition task may be approached differently as a function of input-modality.

1. INTRODUCTION

In the development and evaluation of hearing aids, cochlear implants, and other sensory aids, it is necessary to assess whether or not the device is providing increased access to the acoustic/phonetic patterns of speech. This is most often accomplished through the assessment of speech recognition performance, using percent correct whole-word scoring for lists of open-set monosylllabic words. It is well-known, however, that this approach not only provides information aboutt access to acoustic/phonetic speech patterns, but also about an individual's knowledge of the lexicon.

An alternative approach, which allows for the separation of access to acoustic/phonetic speech patterns from lexical context effects, is the application of simple probability theory to the word recognition task [1]. In isolated word recognition tasks, this involves the relationship between the recognition probabilities of wholes (P_w) to the probabilities for the parts (P_p), which yields a value "j". J indicates the number of independent perceptual units necessary for correct word recognition.

An underlying assumption of "j" is that it is a constant, i.e., that it is not affected by the underlying recognition probabilities for the parts. The purpose of this study was to test this assumption through varying the type and availability of sensory information in two ways: (1) through presentation of stimuli via three input modalities; and, (2) through the use of multiple speakers.

2. METHODS

2.1. Participants

Three male and 21 female young adults (ages 20 to 34 years old) with normal hearing and vision participated in the study.

2.2. Stimuli

The video laser disc recording the the AB short-isophonemic word lists [2] were used to obtain phoneme and whole-word scores. Each list contains 10 CVC words and there are 15 lists in all. Each of the lists were spoken by two male and two female adults with General American dialects. The head and shoulders of each speaker are shown in the recording.

2.3. Instrumentation

The Computer Assisted Speech Perception Testing/Training System developed at the Center for Speech & Hearing Research at the Graduate School of the City University of New York, illustrated in Figure 1, used for stimulus presentation and the scoring of responses. During testing participants were seated four feet from the video monitor. The audio signal was mixed with speech-shaped noise and presented binaurally through TDH-49 headphones, mounted in MX-41 AR cushions. Stimuli were presented at a 0 dB signal-to-noise ratio at a comfortable listening level for each participant.

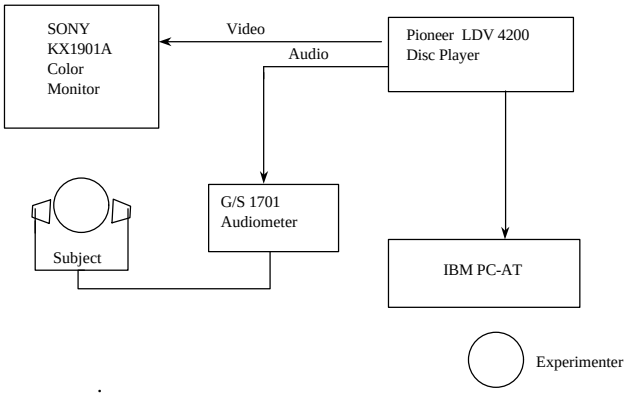

Figure 1. Instrumentation

2.3. Procedure

The participants were tested individually, in four, one-hour sessions. In each session 15 lists spoken by one of the four talkers were presented in all three input-modalities. The order of presentation of the lists was randomized across speakers for

each participant and the order of input-modality was counterbalanced across participants and test sessions. Each test item was presented once and the participants were instructed to repeat what they heard and/or saw. The experimenter then keyed the phonemes correctly perceived into the computer for subsequent scoring and analysis.

3. Results

For each participant, mean percent correct whole word recognition and mean percent correct phoneme recognition scores were obtained, as a function of speaker and input-modality, by averaging performance across the 15 AB-word lists. These data were examined to: (1) confirm the assumption that performance, as measured by percent correct scores, would differ for wholes (i.e., word recognition) and the parts (i.e., phoneme recognition), for each speaker in each input-modality; (2) confirm the assumption that performance, as measured by percent correct word or phoneme recognition scores would differ as a function of speaker and/or modality; and, (3) examine the assumption that *j*-values are not influenced by the underlying recognition probabilities for the parts.

3.1. Performance as measured by word and phoneme recognition percent correct scores for each speaker in each modality

The arcsine squareroot transforms of the percent correct word and percent correct phoneme data were subjected to an analysis of variance (ANOVA) for repeated measures. While the results revealed significant main effects for both modality and linguistic context level (i.e., words vs. phonemes), the most important finding was that for each speaker, there was a significant interaction between modality and linguistic context level. These interactions are illustrated in Figures 2a-d. The results of a Tukey posthoc test on these data revealed that, as expected, for each speaker in each modality, percent correct phoneme recognition was significantly higher than percent correct word recognition.

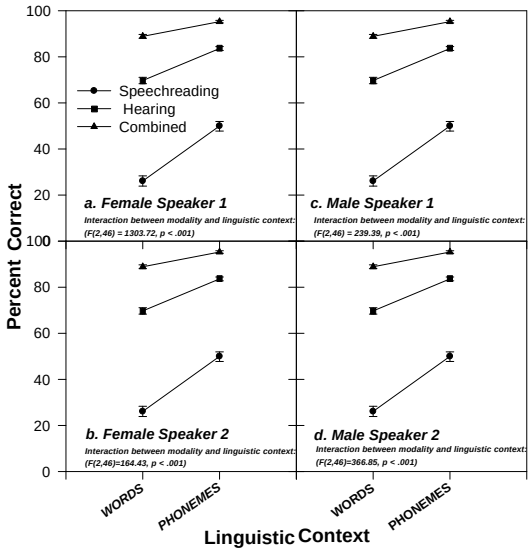

Figure 2a-d. Mean percent correct word and phoneme recognition (+/- 1 SE) for each speaker in each modality.

3.2. Word and phoneme recognition performance as a function of speaker and modality

Figure 3.a illustrates the interaction between speaker and modality for word recognition performance. The results of a repeated measures ANOVA on the arcsine squareroot transforms of the percent correct data revealed significant main effects of speaker (F (3,69) = 34.20, $p < .001$) and modality (F(2,46) = 510.92, $p < .001$), as well as a significant interaction between the two (F(6,138) = 246.66, $p < .001$). The interaction between speaker and modality for phoneme recognition is shown in Figure 3b. The results of a repeated measures ANOVA on these data also revealed significant main effects of speaker (F (3,69) = 6.94, $p < .001$), modality (F (2,46) = 443.87, $p < .001$), and their interaction (F (6,138) = 78.56, $p < .001$). These results were taken to confirm the assumption that the manipulations of speaker and modality did provide differing recognition probabilities for the parts.

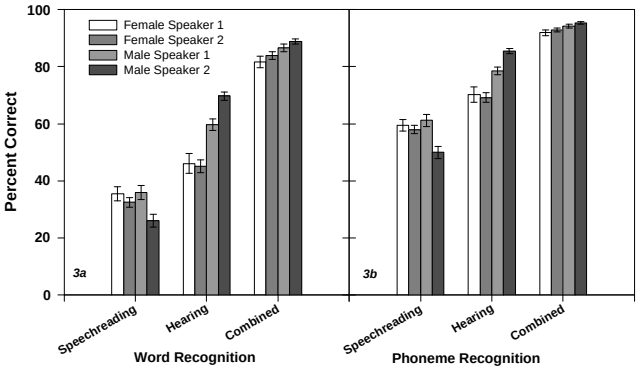

Figure 3a-b. Mean percent correct scores (+/- 1 SE) obtained under each input modality for each speaker.

3.3. Effect of speaker and modality on *j*-values

The primary goal of this study was to examine the assumption that *j*-values are independent of underlying recognition probabilities for the parts. To examine this hypothesis, the percent correct word whole word and phoneme data were used to generate *j*-values for each participant, as a function of speaker and input-modality, using the formula:

$$j = \log (P_w)/\log (P_p) \qquad (1)$$

where P_w is the probability of recognition of words and P_p is the probability of recognition of phonemes. The resultant *j*-values were then subjected to a repeated measures ANOVA.

Figure 4 illustrated the main effect of speaker. As hypothesized, *j*-values were independent of the underlying recognition properties for the speakers ($F(1,69) = 1.02$, $p = .39$)

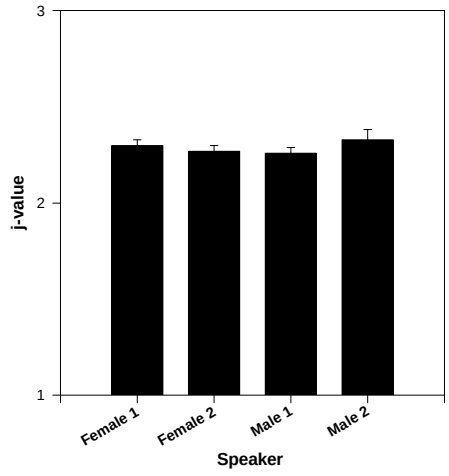

Figure 4. Mean *j*-values (+/- 1 SE) as a function of speaker.

The main effect of modality on *j*-values is illustrated in Figure 5. Contrary to expectations, there were significant differences in *j*-values between the three input-modalities ($F(2,46) = 73.39$, $p < .001$).

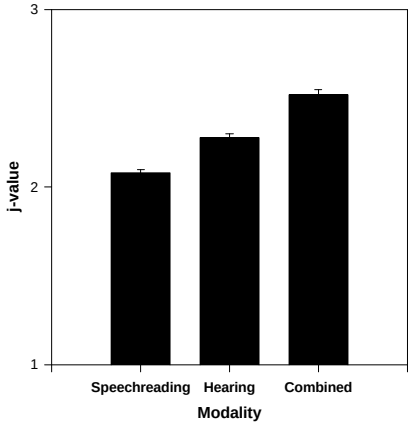

Figure 5. Mean *j*-values (+/- 1 SE) as a function of modality.

The interaction between speaker and modality, as shown in Figure 6, was also found to be significant ($F(6,138) = 3.67$, p , .002). Post hoc analysis, using the Tukey tesk, revealed that within each modality, the differences between speakers were not reliably different. However, for each speaker, the differences in *j*-values for each input-modality were significant.

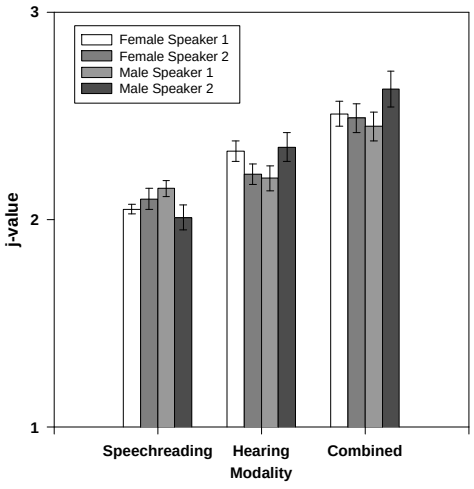

Figure 6. Mean *j*-values for speakers (+/- 1 SE) as a function of modality.

4. Summary and Conclusions

The measurement of speech recognition using lists of open-set monosyllabic words involves assessment of both access to the acoustic/phonetic patterns of speech and an individual's knowledge of the lexicon. This was confirmed by comparison of percent correct word and phoneme recognition scores across speakers and input-modalities. In developing and evaluating sensory aids there is a need to separate lexical knowledge from access to acoustic/phonetic speech patterns. It is proposed that this can be accomplished through the application of simple probability theory [1]. An underlying assumption of this approach, however, is that the available sensory data in the signal, as estimated from the underlying recognition probabilities for phonemes, will not influence the use of lexical context. This assumption was examined by manipulating both speakers and input-modalities in a standard word recognition task. The finding of significant speaker and modality effects on percent correct phoneme and word recognition supports the assumption that these factors can be used to manipulate the availability of sensory data. When this was done, *j*-values, indices of the use of lexical context, were found to be independent of the manipulation of speakers, both collapsed across and within modalities. They were not, however, independent of modality. These findings were interpreted to suggest that individuals may be using different processing strategies when approaching the speech recognition task via speechreading alone, hearing alone, and bi-modally.

REFERENCES

[1] Boothroyd, A. and Nittrouer, S. 1988. Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America, 84*, 101-114.

[2] Boothroyd, A. 1968. Developments in speech audiometry. *Sound, 2*, 3-10.