MODALITY, PERCEPTUAL ENCODING SPEED, AND TIME-COURSE OF PHONETIC INFORMATION

Philip Franz Seitz and Ken W. Grant

Army Audiology and Speech Center
Walter Reed Army Medical Center
Washington, DC 20307-5001

ABSTRACT

This study examined the perceptual processing of time-gated auditory-visual (AV), auditory (A), and visual (V) spoken words. The primary goal was to assess the extent to which stimulus information versus perceptual processing limitations underlie modality-related perceptual encoding speed differences in AV, A, and V spoken word recognition. Another goal was to add to the scant literature on the comparative time-course of phonetic information in AV, A, and V spoken words [1]. In terms of duration of speech signal required for accurate word identification, it was found that AV<A<V. For individual word stimuli, there were strong predictive relations between unimodal encoding speed and gating measures. Perceptual encoding of V words is slower than predicted based on stimulus information alone.

1. INTRODUCTION

A slowing of perceptual encoding due to combined effects of aging and poor stimulus clarity is one possible factor underlying the deficits in comprehension of fluent speech shown by many older people with hearing loss [2, 3]. Similarly, a component of the comprehension advantage afforded by AV over A speech might be that perceptual encoding operates more efficiently on AV inputs, releasing attentional resources for cognitive comprehension processes [4]. Indeed, an earlier experiment demonstrated significant modality-related perceptual encoding speed differences in a group of 26 older subjects with acquired hearing loss [5]. In that experiment, subjects performed the Sternberg memory scanning task in the three modalities using an ensemble of 10 spoken words that they could recognize with near perfect accuracy in all modalities. Least-squares linear models were fit to the memory set size × reaction time (RT) group data for AV, A, and V conditions. The models' y-intercepts represented perceptual encoding speed plus residual latency: RT_{AV} = 703 msec, RT_A = 763 msec, and RT_V = 867 msec.

Using a time-gated word identification task ("gating"), the present experiment assessed the extent to which these modality-related RT differences can be accounted for by differences in the amount or time-course of phonetic information afforded by the modalities, versus possible differences in the structure of perceptual processing associated with modality.

The *gating paradigm* investigates the time-course of spoken word recognition by presenting words repeatedly, starting with a short portion ("gate") including the part of the signal determined to be the word's onset, and progressively increasing the duration (number of gates) from the onset [6, 7]. Subjects guess the identity of the word on each presentation and make a confidence rating for the guess. The duration at which a subject correctly guesses the word and does not subsequently change his guess is defined as the isolation point (IP). The confidence ratings support the concept of a total acceptance point (TAP), defined as the stimulus duration at which a subject correctly identifies the stimulus and gives it a high confidence rating without subsequently changing identification or lowering confidence rating.

The gating paradigm generally has been employed for studying properties of the mental lexicon. The present experiment used it to investigate the time-course of phonetic information availability, as a function of modality, for the closed set of ten spoken words on which perceptual encoding speed measures were obtained in the earlier memory scanning experiment [5]. For encoding speed, it is obvious that, all other factors being equal, a word with an earlier IP or TAP should show an RT advantage over a word with a later IP or TAP. (Because the decision to respond is affected by an

internal criterion, or fixation of belief, as well as by stimulus information, the TAP might be more predictive of RT than the IP.) Likewise, a modality in which words' IPs or TAPs are early on the average should support faster perceptual encoding than a modality in which they are later, to the extent that encoding speed is a function of stimulus information [8]. If phonetic information differences account for modality-related encoding speed differences, it would support the idea that a general data limitation underlies modality effects on RT to spoken words [9]. On the other hand, if modality-related encoding speed differences are greater than predicted from phonetic information measures, then it is reasonable to suppose that some aspect of perceptual processing, that is, a resource limitation, is responsible for part of the encoding speed differences. This reasoning depends on the existence of a predictive relationship between encoding speed and IP or TAP, which will be demonstrated below.

2. METHODS

2.1 Subjects

Twenty-four older people with acquired mild-to-moderate sensorineural hearing loss participated in the experiment. Their mean age was 66 y (sd = 5.6) and their mean better-ear 3-frequency (.5, 1, 2 kHz) pure-tone average audiometric threshold was 37 dB HL (sd = 11.6). All had participated in the earlier memory scanning experiment [5].

2.2 Stimuli

There were 20 stimulus words consisting of two tokens of each of 10 monosyllabic words from the CID W-22 word list: *bread*, *pie*, *live* (/lɪv/), *jump*, *felt*, *three*, *wool*, *star*, *ears*, *owl*. The words were spoken carefully by one male talker, video recorded on SVHS tape, and dubbed onto an optical disc (Panasonic TQ-FH331, Panasonic LQ-3031/2T recorder/player). The words were selected on the basis of pilot testing in order to allow perfectly accurate closed set identification by lipreading.

The stimulus durations were defined as encompassing the first and last video frames (1/30 sec frame duration) in which a word's audio signal was observable, and only these frames were dubbed onto the optical disc. Stimuli were chosen from among many available tokens so that all had the same duration, 20 frames (666.6 msec). A single silent video frame showing the talker in a resting position was dubbed to precede each of the 20 stimuli. This frame, which was identical for all stimuli, was displayed for 400 msec before starting video playback in order to make the stimuli appear more natural.

2.3 Procedure

The experiment required one 150-minute session per subject. All subjects passed a 20/30 (corrected) vision screening test using a Snellen chart and were tested individually in an audiometric test booth. They sat at a table equipped with a touchscreen terminal, facing a 20-inch color video monitor positioned six feet away at eye level. Auditory signals were routed from the optical disc player through an audiometer to an insert earphone. Auditory stimuli were presented monotically to the better ear at each subject's most comfortable level.

Subjects were given a written list of the 10 stimulus words to which they could refer during testing. Before beginning the time-gated word identification task, all 24 subjects demonstrated 100% accurate identification of the 20 complete, intact stimuli in all three modalities by making 20/20 correct verbal responses in three consecutive scrambled-order blocks of trials.

Subjects performed the time-gated word identification task in three parts, one part per modality (AV, A, V). Four subjects were assigned to each of the six possible modality condition orders.

In an experimental trial, a portion of a stimulus word was presented, including the word's onset. On the touchscreen, the subject touched one of ten orthographic words to indicate the one he thought was presented, then touched a point along a horizontal line to indicate confidence in the identification, and finally touched a "confirm" label to begin the next trial. No feedback was given. Subjects were instructed to guess if they did not know which word was presented. The confidence line was labeled "not sure" at its left end and "very sure" at its right end. It was explained to subjects as representing a range from 0% to 100% confidence in the word identification response. Confidence responses were discretized into 30 intervals.

The time-gated word recognition task was implemented in the "duration blocked" format [7]. In this format, all stimuli are presented in scrambled

order at a given duration, then are re-scrambled and presented again at the next longer duration. Subjects performed the task in each modality until they met the joint criteria of an average of at least 95% correct and at least 93% confidence for the 20 words at a given duration, or until all 20 gates (the words' entire duration) had been presented.

3. RESULTS

For each subject in each modality, the IP and TAP measures were computed separately for each of the 10 words, averaging the two tokens per word. The IP for a given word was computed as the duration at which subjects correctly identified both tokens of the word and did not subsequently mis-identify either token. The TAP was computed as the duration at which subjects correctly identified both tokens of the word with an average confidence over the tokens of at least 80% and never subsequently mis-identified either token or had a lower average confidence rating for them. Overall IP and TAP measures for individual subjects were computed for each modality as the averages of the measures for the ten words.

Figure 1 shows IP and TAP averages across subjects and words for the three modalities as a function of gate, that is, stimulus duration presented.

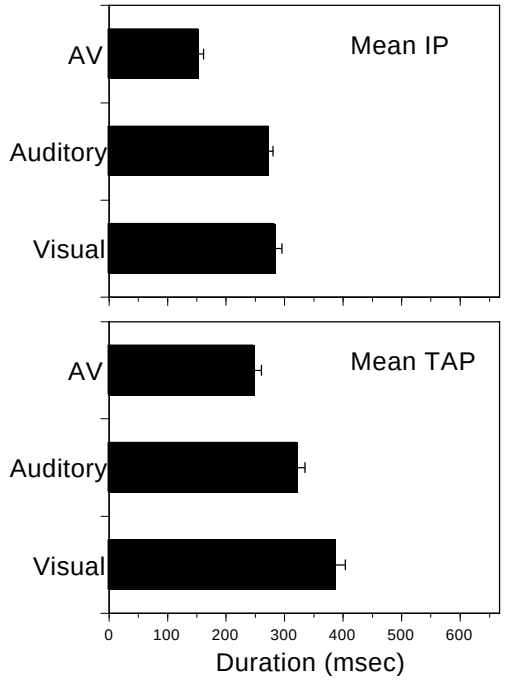

Figure 1: Averages over 24 subjects and 10 words (2 tokens per word) of words' IP and TAP for three modalities. Note that the end of the x-axis (666 msec) represents stimulus words' total duration. Error bars represent standard error of the mean.

It is clear that more information is available earlier from AV words than from A or V words by both the IP and TAP measures, but the story for differences between A and V is more complex. Repeated-measures ANOVAs with two within-subjects factors (Modality, Word) were run separately on the IP and TAP data using subject as the random variable. Results indicated a significant main effect of Modality for both IP [$F(2,46)=81.98$, $p<.001$] and TAP [$F(2,46)=22.7$, $p<.001$]. Post-hoc paired-samples t tests showed that the difference between IP_A and IP_V was not significant [$t(1,23)=0.76$, *ns*], but that the differences between IP_A and IP_{AV} [$t(1,23)=12.5$, $p<.001$] and IP_V and IP_{AV} [$t(1,23)=12.0$, $p<.001$] were. The situation for the TAP measure was different: TAP_A was significantly earlier than TAP_V [$t(1,23)=3.5$, $p<.002$], and TAP_{AV} was significantly earlier than TAP_A [$t(1,23)=4.8$, $p<.001$] and TAP_V [$t(1,23)=11.1$, $p<.001$].

Figure 2 illustrates modality-related differences in the time-course of phonetic information availability reflected in a typical subject's performance. Although this subject's IP_A and IP_V averaged across words are nearly identical (250 msec and 257 msec,

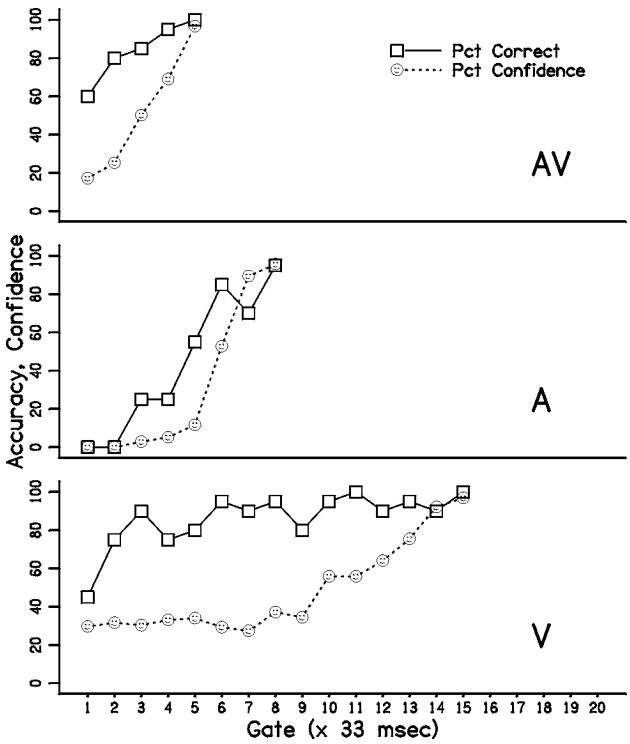

Figure 2: Time-gated word recognition data for one subject: Percent correct and percent confidence averaged over 10 words (2 tokens per word) in three modalities.

respectively), he received much more information in V over the first several gates than in A. On the other hand, once information started to become available in A, by around gate 6, his confidence as well as accuracy increased rapidly. Confidence in V remained low despite the fact that 95% correct identification was achieved in V by gate 6. These data are consistent with Smeele's gating study of AV, A, and V nonsense syllables, which reported that place of articulation features for consonants were available earlier in V than in A [1].

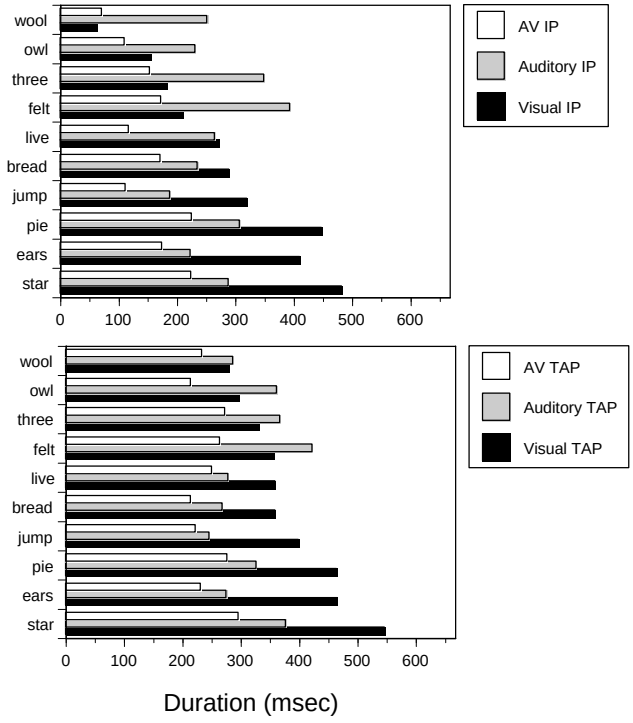

Figure 3: Isolation Points (IPs) and Total Acceptance Points (TAPs) for 10 words (2 tokens per word) in three modalities, averaged over 24 subjects. Note that words are arranged in ascending order of Visual TAP.

There were large word-related differences in IP and TAP in each of the modalities, as shown in Figure 3. The main effect of Word was significant for both IP [$F(9,207)=24.8$, $p<.001$] and TAP [$F(9,207)=7.7$, $p<.001$]. The interaction of Word and Modality was also significant for both measures [IP: $F(18,414)=16.0$, $p<.001$; TAP: $F(18,414)=4.2$, $p<.001$], indicating that the patterns of word-related differences in IP and TAP were not the same in the three modalities. Post-hoc tests of contrasts between A and V indicated that the IP_V of *felt*, *three*, *wool*, and *owl* were significantly earlier than their IP_A; The IP_A of *pie*, *jump*, *star* and *ears* were significantly earlier than their IP_V; and *bread* and *live* had non-significantly different IP_A and IP_V. The four words having early IP_V had TAP_A and TAP_V that were not significantly different, whereas the other six words all had TAP_A that were significantly earlier than their TAP_V.

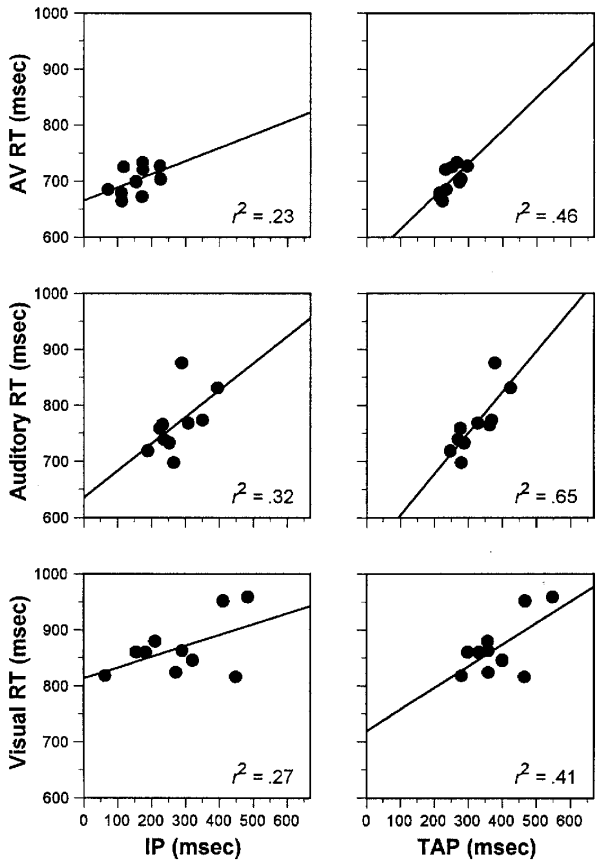

Figure 4: Bivariate distributions of stimulus information measures (IP, TAP) and perceptual encoding speed (RT) for 10 words (2 tokens per word), averaged over 24 subjects, with coefficients of determination.

In order to examine the relationship of the gating measures to perceptual encoding speed, subjects' RT data from the earlier memory scanning experiment were averaged over trials in which each of the 10 words was the probe [5]. The IP and TAP for each word also were computed by averaging over the 24 subjects. The six panels of Figure 4 illustrate the predictive relations between the time-course of stimulus information, as measured by gating, and perceptual encoding speed. In all three

modalities, TAP explains more of the variance in encoding speed than does IP, and all Pearson correlations between TAP and RT are significant (α=.05, 1-tailed). IP is only significantly correlated with RT_A.

4. DISCUSSION

In light of the relationship between TAP and encoding speed, the finding that TAP_V is significantly later than TAP_A and TAP_{AV} seems to support the idea that modality-related encoding speed differences are actually not due to modality *per se*, but rather to stimulus information more generally. However, a closer comparison of the TAP and RT data leaves open the possibility that modality does affect the structure of perceptual processing. Such a possibility is suggested by considering what the differences among modalities' encoding speeds would be if they were entirely predicted by TAP differences. The average TAP_A-TAP_{AV} difference of 73 msec is 13 msec greater than the 60 msec encoding speed difference between A and AV. That is, the TAP_A-TAP_{AV} difference slightly overpredicts the A-AV difference in encoding speed; AV encoding is not quite as fast, compared to A, as stimulus information would predict. On the other hand, the TAP_V-TAP_{AV} difference of 138 msec is 26 msec *less* than the V-AV encoding speed difference, and the TAP_V-TAP_A difference of 65 msec is 39 msec *less* than the V-A encoding speed difference. That is, the TAP_V-TAP_{AV} and TAP_V-TAP_A differences underpredict the V-AV and V-A encoding speed differences; V is even slower, compared to AV and A, than stimulus information would predict. Although stimulus information accounts for a significant portion of RT variance, there appears to be an additional amount of time added to RT when a spoken word stimulus does not include auditory information. This pattern is consistent with the notion of an additional processing stage involved in recognition of V words, in which they are translated into an auditory representation in order to make contact with memory representations [10]. A literal interpretation of the present findings is that the additional processing required by V input takes about 30 msec. Recent physiological work on evoked responses associated with perceptual encoding and integration has the potential to corroborate behavioral findings on the time-course of perceptual processing [11].

5. ACKNOWLEDGMENTS

This work was supported by research grant numbers R29 DC01643 and R29 DC00792 and by Research Supplement for Underrepresented Minorities R29 DC01643-02/3S1 from the National Institutes of Health, National Institute on Deafness and Other Communication Disorders. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or reflecting the views of the Department of the Army or the Department of Defense.

6. REFERENCES

1. Smeele, P.M.T., *Perceiving Speech: Integrating Auditory and Visual Speech*, Technische Universiteit Delft doctoral dissertation, 1994.
2. Working Group on Speech Understanding and Aging, "Speech understanding and aging," *Journal of the Acoustical Society of America*, *83*, 859-895, 1988.
3. Gordon-Salant, S., and Fitzgibbons, P.J., "Selected cognitive factors and speech recognition performance among young and elderly listeners," *Journal of Speech, Language, and Hearing Research*, *40*, 423-431, 1997.
4. Reisberg, D., McLean, J., and Goldfield, A., "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli," in B. Dodd and R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading* (pp. 97-113). Hillsdale, NJ: Lawrence Erlbaum Assoc., 1987.
5. Seitz, P.F., "Hearing-impaired perceivers' encoding and retrieval speeds for auditory, visual, and audiovisual spoken words," *Journal of the Acoustical Society of America*, *101*, 3155, 1997.
6. Grosjean, F., "Spoken word recognition processes and the gating paradigm," *Perception & Psychophysics*, *28*, 267-283, 1980.
7. Walley, A.C., Michela, V.L., and Wood, D.R., "The gating paradigm: Effects of presentation format on spoken word recognition by children and adults," *Perception & Psychophysics*, *57*, 343-351, 1995.
8. Massaro, D.W., *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press, 1998.
9. Norman, D.A., and Bobrow, D.G., "On data-limited and resource-limited processes," *Cognitive Psychology*, *7*, 44-64, 1975.
10. Lyxell, B., Rönnberg, J., and Samuelsson, S., "Internal speech functioning and speechreading in deafened and normal hearing adults," *Scandinavian Audiology*, *23*, 179-185, 1994.
11. Levänen, S., "Neuromagnetic studies of human auditory cortex function and reorganization," *Scandinavian Audiology*, *27* (Suppl. 49), 1-6, 1998.