OSU Working Papers in Linguistics 53, 168 - 199

Turning Phonology Inside Out, or Testing the Relative Salience of Audio-Visual Cues for Place of Articulation*

Steve Winters
swinters@ling.ohio-state.edu

Jun (1995) and Hume (1998) motivate phonological analyses of cross-linguistic trends in place assimilation and consonant/consonant metathesis by appealing to putatively universal rankings of the perceptual salience of stop place cues. Experimental support for such salience rankings is sparse, perhaps because of the difficulty of eliminating "inside-out" effects in establishing the *inherent* salience of phonetic cues. This study attempted to explicitly test speculative claims about cue salience for stop place using audio-visual stimuli in an experimental paradigm that minimized the "inside-out" effects of linguistic structure on speech perception. Salience was gauged by evaluating the perceptual effects of adding acoustic or visual information to experimental stimuli. Results showed that labials have the most salient place cues in either the auditory or visual modality, contrary to what some theoreticians would have predicted. However, dorsals gain the most salience from adding acoustic information to the signal, suggesting that perhaps only acoustic cues have "outside-in" effects on phonological structure.

INTRODUCTION

Recent work in Optimality Theory has suggested that certain cross-linguistic phonological processes may be based on aspects of speech perception. Jun (1995), for instance, proposes a meta "preservation" constraint of the following form:

(1) Pres(X(Y)): Preserve perceptual cues for X (place or manner of articulation) of Y (a segmental class)

* Thanks go to Keith Johnson, Beth Hume, Mary Beckman, Jenny Vannest, Kiyoko Yoneyama, Matt Makashay, Janice Fon, Amalia Arvaniti and all of the linguistics lab's Psychoids for helpful comments and criticisms on this paper. I am also grateful to Steve Hartman Keiser and Rebecca Herman for their help in the production of the stimuli used in this experiment. This material is based upon work supported under a National Science Foundation Graduate Fellowship.

Furthermore, Jun proposes that all preservation constraints for some place or manner of articulation are universally ranked with respect to each other:

(2) Universal ranking: Pres(M(N)) >> Pres(M(R)), where N's acoustic cues are stronger than R's cues for M.

For instance, some of the universal rankings for place preservation include:

(3) Pres(pl(dor¯)) >> Pres(pl(lab¯)) >> Pres(pl(cor¯))

(4) Pres(pl(onset) >> Pres(pl(coda))

(5) Pres(pl(stops) >> Pres(pl(nasals))

Jun provides the following example (among others) of how such universal rankings of constraints might interact with the articulatory WEAKENING constraint to account for place assimilation in Korean:

(6) Example (Korean): /ip + ko/ → [ikko] 'wear and...'

	/ip + ko/	Pres (pl(onset))	WEAKENING	Pres(pl(coda))
	ipko		**!	
☺	ikko		*	*
	ippo	*!	*	

WEAKENING is a constraint that prohibits consonantal articulations; candidates [ippo] or [ikko] are preferable to input [ipko] since they both have only one consonantal articulation as opposed to two. In general, place assimilation will occur in a language according to how it ranks WEAKENING with respect to the various preservation constraints.

This phonological account of assimilation is a formal treatment of Kohler's (1990) production hypothesis, which states that speakers make more effort to produce stronger acoustic cues than weaker ones. The motivation for these formal structures comes from two distinct sources, the first of which is cross-linguistic patterns of place assimilation. Jun examines phonological processes in a number of languages and notes that in none of them do nasals assimilate while stops do not. In terms of place of articulation, dorsals in

coda position would never assimilate unless both labials and coronals do as well. The rankings of preservation constraints thus read like an implicational hierarchy for place assimilation-- labials will assimilate only if coronals do as well, and so on and so forth.

The universal rankings of preservation constraints might, therefore, be completely justified from a strictly phonological point of view, but Jun goes one step further and attempts to motivate them with phonetic facts about the perceptual salience of place cues. After all, the rankings of preservation constraints should fall out from which cues are "stronger" than others (according to (2)). So, Jun bases his ranking of preservation of cues for place in coda position by appealing to a number of speculative claims about which of these cues are more salient than others. Coronals are at the bottom of the list because their transitions are shorter and have relatively "small excursions" when compared to dorsals or labials. Dorsals, in turn, have stronger cues than labials because of the supposed acoustic prominence of the "velar pinch" (as noted in Stevens (1989)).

Hume (1998) makes similar appeals to the perceptual salience of place cues in order to motivate her phonological analysis of consonant/consonant metathesis. Hume notes that labials have a unique cross-linguistic tendency to undergo this unusual process. From a phonological perspective alone, then, there is reason to believe that labials are somehow special among the various stop consonants, but Hume tries to back up this claim further by appealing to the "perceptual vulnerability" of labials. Hume does not elevate "perceptual vulnerability" to the formal status of a meta-constraint, but she does use it as the phonetic background for the relative ranking of the specific constraints that drive consonant/consonant metathesis involving labials. For instance, Hume proposes that the weak release bursts of labials do not add much to their perceptual salience (Ohala (1990)), and so it would be preferable to place them in coda position as opposed to onset position. Place cues are also (presumably) more salient in stressed syllables than in unstressed syllables. These two facts about perceptual salience together motivate the ranking of *labial/C-V >> *labial/V-C, as Hume proposes is the case in Kui, where labials undergo consonant/consonant metathesis into stressed coda position.

(7) Example (Kui): /ag + ba/ → [ábga] 'to be fitting'

	/ag + ba/	*labial/C-V	*labial/V-C	LINEARITY
☺	ábga		*	
	ágba	*!		

Hume and Jun's analyses both refer to the perceptual "salience" of certain stop places of articulation in order to account for why metathesis and assimilation occur; interestingly, though, the two phonologists reach different conclusions about which places of articulation are more or less salient than others. In the ranking in (3), Jun proposes that dorsals are most salient, followed by labials and then coronals. For Hume, however, labials have the least salient place cues.

The fact that these phonologists come to different conclusions about the relative strength of cues for different places of articulation is no accident. Both phonologists were able to justify their analyses with claims made by different speech researchers, who should be authorities on what may or may not constitute a strong acoustic cue for a place of articulation. However, speech perception researchers have not been able to establish which stop place cues are stronger or weaker than others. This empirical question remains unanswered despite the best efforts of studies such as Miller and Nicely (1955), Malécot (1958) and Wang and Bilger (1973) (among others), which have all tried to solve this problem but yielded conflicting and inconclusive results using experimental paradigms of varying design and purpose.

Miller and Nicely (1955) presented listeners with 16 different consonant sounds (including stops, fricatives, nasals, voiced and voiceless sounds) in various levels of noise, and asked the listeners to identify them in an open-response format. From the resultant 16x16 confusion matrices, it is possible to pull out the stops and determine (using the "I" sensitivity measure described below) that their listeners found coronals more salient than both dorsals and labials, neither of which differed significantly from each other. Wang and Bilger (1973) used a similar paradigm, although they threw affricates into the consonantal mix, added productions with the vowels /i/ and /u/, put consonants in both onset and coda position, and also used sound level reduction in addition to introducing noise into the signal. Their results showed that labials and coronals were equally salient in the onset condition—and both were more salient than dorsals—and, in the coda condition, coronals were more salient than labials which, in turn, were more salient than dorsals. Malécot(1958) took a completely different tack and experimented with adding or removing bursts and transitions from stop consonants in

coda position. His results showed that labials were more salient than both coronals and dorsals, which did not differ significantly from each other.

Others have approached the problem of measuring place cue salience in different ways but each new attempt seems to confuse the situation more than it does to help clarify it. One good reason for this confusion may be that it is simply so difficult to determine what is *inherently* 'salient' or 'strong' about an acoustic cue as opposed to what listeners might project onto the speech signal in developing a linguistic interpretation of it. Such interpretive projections in speech perception are commonly called "top-down effects," as an extension of the metaphor that certain levels of linguistic structure are "higher up" than others. For example, a listener's upper-level semantic, pragmatic and syntactic knowledge might enable them to perceive the word "nine" before they have heard little (if any) acoustic input for that word in the following sentence (Lieberman (1963)):

(8) A stitch in time saves nine.

Top-down information has similar influences within single-word contexts as well; Warren (1970) showed that replacing the fricative /s/ with a non-linguistic noise such as a cough in a word like 'legislation' has little or no effect on listeners' perception of that word. Many listeners did not even hear the cough (as such) at all, and most of those who did interpreted it as occurring sometime after the word had ended. Phonological effects on speech perception should be familiar to anyone who has ever attempted to learn phonetic transcription; most native-English speakers hear initial /tl-/ clusters as [kl], since such clusters are not permitted by English phonology. Precisely the opposite is true of Navajo speakers, who interpret an English word like 'clock' as /tlak/. (Schaengold, 1999) Any attempt to objectively establish the inherent salience of some acoustic cue would have to eliminate the possibility of any of these top-down influences intruding in upon the perceptual task. Since speech perception science has not yet finished experimentally testing the myriad possibilities of top-down influences that may exist in perception, it is difficult to claim for certain in any experimental paradigm that such influences have been eliminated completely.

There is another way of thinking about top-down influences on speech perception; since they essentially consist of mental structures that a listener imposes on an incoming speech signal, one might think of them as "inside-out" processes. That is, they transform linguistic structures inside the mind into perceived physical realities in the outside world. Analogously, if external cues for some linguistic structure have a role in motivating some universal phonological constraint, they can only do so by virtue of what I would like to call "outside-in" effects in speech perception. These could be characterized as the internalization of perceivable structures in the external speech signal as linguistic (or phonological) structures inside the human mind.

In modern linguistics, phonologists have generally been interested in inside-out processes. In other words, most phonological analyses would hope to explain how the mind influences the patterns of sounds used in language as opposed to the other way around. With their emphasis on possible outside-in influences on phonological structures, though, the optimality theoretic analyses of Jun and Hume (among others) seem to represent a new trend in doing phonology. In Optimality Theory, phonologists are not simply content to characterize what phonological processes *may* happen in language; they want to understand and formalize *why* certain processes happen and others do not. Though cognitive coherence and simplicity may be the most fundamental force in shaping linguistic structures, most optimality theoreticians would concede that the communicative efficacy of sound structures plays an important role as well. Such theoreticians would not, therefore, strictly relegate phonological phenomena to an internal role in the mind but recognize that it has externally-based features as well, due to a language user's need to perceive as well as produce the phonological structures of their language.

This theoretical strategy can provide plenty of work for speech perception researchers even though it may unnecessarily complicate the world of phonological theory. Outside-in effects more easily submit to experimental verification than to introspective analysis (a linguist's usual scientific tool of choice). Objects and events in the external world can be manipulated and reproduced with relative ease, while it is almost impossible for an experimenter to manipulate or reproduce the internal structures of the human mind. Thus, an experimental test of the outside-in effects of speech stimuli could simply involve the presentation of such stimuli to listeners who would be asked to

categorize them in terms of some phonological structure. Since phonological structure is a necessary outcome of any speech perception task, completely eliminating the possibility of any inside-out influences in such an experiment is impossible. An experimenter could, at least, minimize the other "inside-out" influences by extracting the stimulus from any pragmatic or syntactic context and maximizing the use of "nonsense" words to avoid word-level semantic effects. An experiment of this kind could provide one empirical method of verifying what outside-in effects may exist in language (as well as their relative strengths).

Performing an experiment of this kind would also be an appropriate test of the validity of Jun's and Hume's claims about the relative salience of cues for place of articulation. There is considerable evidence, however, that empirically testing the salience of place cues--and thereby resolving the discrepancy between Jun's and Hume's salience rankings--would have to involve a perceptual experiment that used audio-visual stimuli. Many speech perception studies have shown that listeners perceive place not only through acoustic cues such as bursts or transitions, but also through visual cues, such as movements of the lips, tongue or jaw. One of the most well-known of these visual perception studies is McGurk and MacDonald (1976), in which it was shown that people's perception of audio-visually mismatched stimuli can change depending on which place of articulation is presented auditorily and which is presented visually. Some basic examples of how this phenomenon works include:

(9) Typical McGurk effects

Subject	sees:	+	hears:	⇒	perceives:
	ba	+	ga	⇒	ba
	ga	+	ba	⇒	da or bga or gba

This bizarre phenomenon significantly changes our basic understanding of speech perception not only because it incontrovertibly shows (as others have shown) that people use visual information in perceiving speech, but also that people sometimes attach more perceptual importance to visual information than to acoustic cues. The McGurk effect is especially strong in stop consonants, which have minimal acoustic cues but

comparatively noticeable visual cues for place of articulation. Jun and Hume ignored the comparative importance of visual cues in establishing hierarchies of stop place salience; they only considered the strength of a stop's acoustic cues. Whether or not visual cues for stop place motivate the phonological proposals of Jun and Hume is unclear; but it *is* certain that visual cues can contribute to the perceptual salience of stop place. An "outside-in" experimental paradigm could determine the *inherent* salience of any visual or acoustic cue for stop place and thereby determine whether the theoretical proposals of Jun and Hume actually correspond to the empirical reality of stop place salience. According to Jun, for instance, one would expect dorsals to have the most salient place cues; according to Hume, however, one would expect labials to be the least salient. In considering visual perception studies, though, one would expect labials to be the *most* salient. And yet other possibilities exist, too: coronals might be the most salient, for instance. Which one of these possibilities reflects empirical reality is unknown, however, and therefore any phonological claims that are based on assumptions about place salience remain untested conjectures.

This study attempted to explicitly test such speculative claims about the salience of stop place by using audio-visual stimuli in an experimental paradigm that minimized the "inside-out" effects of linguistic structure on speech perception. The results of this experiment could hopefully not only improve the current understanding of stop place perception but also provide the necessary empirical framework for phonological analyses that appeal to perceptual facts for motivation.

METHOD

In attempting to gauge the relative strength of cues for place of articulation, this study adopted a strategy of comparative analysis: it compared listeners' success rates at perceiving place when they were presented with normal phonetic information as opposed to little or no information. For instance, listeners heard or saw identical stimuli with both normal acoustic information and minimal acoustic information. The salience of an acoustic cue for a particular stop place, then, was considered to be how much it contributed to a listener's perception success when it was added to the minimally informative signal. The salience of visual cues, on the other hand, would correspond to

how much they helped listener performance between conditions with normal visual information and no visual information. What changes had been made in the external speech signal between the two conditions could therefore be held responsible for the changes that occurred in the perceiver's comprehension of the signal, and the resultant experimental effects could be considered genuinely "outside-in."

The first step in setting up such comparable experimental conditions was to create audio-visual stimuli for the listeners to try to perceive. Video recordings were made of both a male native speaker and a female native speaker of American English. Each speaker was instructed to read from a script that was placed just underneath the lens of the camera; the speakers sat approximately three to four feet from the camera and were shot from the shoulders up. The video recordings were made inside a sound booth with an 8 mm camcorder. An external microphone hanging from the ceiling of the sound booth above the speaker's head provided the audio portion of the recording.

The script from which the speakers read included stop productions in a variety of phonological contexts. The speakers were asked to produce voiced stops only, in labial, coronal and dorsal places of articulation, as both nasal and oral stops, in both onset and coda position, with both the vowel /a/ and /i/, and in both stressed and unstressed syllables. All of these variations were included to test Jun's and Hume's rankings of salience and preservation constraints. Production with the two different vowels was included to provide a broader and more realistic coarticulatory context and also because it was suspected that visual cues would be stronger when produced with a large jaw opening for /a/ than with the comparatively small opening for /i/. In order to simplify this multi-faceted production task, the speakers read two syllable nonsense words with one stressed and one unstressed syllable, with the same stop at both the beginning and the end of the word. The two syllables were separated with a production of /h/, which was selected because of its lack of potentially confusing place cue information. In short, this meant the speakers had to produce all of the following forms:

(10) Production script

báhab	baháb	máham	mahám
dáhad	dahád	náhan	nahán
gáhag	gahág	ŋáhaŋ	ŋaháŋ
bíhib	bihíb	míhim	mihím
díhid	dihíd	níhin	nihín
gíhig	gihíg	ŋíhiŋ	ŋihíŋ

Each speaker was asked to produce each item on the list at least three times. From the resultant video recording, one of each speaker's productions for each token was selected to become a stimulus in the place perception experiment. These tokens were digitized into 320x240 video clips using Adobe Premiere on Macintosh. Due to a glitch in Premiere's digitization algorithm, the audio and visual portions of the recording had to be aligned manually after each digitization. This was done by digitizing three consecutive tokens at a time and then realigning the audio portion of the recording so that all three tokens of the sequence appeared to be properly aligned. In general this meant delaying the beginning of the audio until the video had already played for six to eight frames (approximately .2 to .3 seconds). Judgments of proper alignment had to be made by the editor's intuition based on video landmarks like lip opening, jaw lowering and acoustic landmarks like vowel offset. Previous research (e.g., Munhall et al. (1996)) indicates that any minor misalignments in the stimuli that may have resulted from this process probably did not affect listener integration of the visual and audio signals.

After the video tokens had been digitized and properly aligned, individual CV or VC tokens were clipped for use as stimuli in the perception experiment (see Figure 1). For CV tokens, the video was cut at the last frame before the onset of frication in the medial /h/ in the original production, and for VC tokens, the video was cut so that it began with the first frame after the offset of frication for the medial /h/. Examinations of the waveform of the video's audio portion along with frame-by-frame playback of the video made it both possible and easy to determine where these audio landmarks occurred on the recording. In addition, CV tokens were cut to begin ten frames (approximately

1/3rd of a second) before the onset of the acoustic waveform on the recording, and VC tokens were cut to end ten frames after the offset of any acoustics on the recording. The inclusion of such pre- and post-acoustic material in the tokens meant that subjects could see visible gestures in the speaker's face before or after they had made any acoustic effects.

After these digital cuts had been made, the first frame of each clip was saved as a .PICT file and expanded as a still picture to make up the entire first second of each clip. Previous experimentation (Strand and Johnson (1996)) has shown that subjects need such preparatory still shots in order to visually orient themselves to a face before they try to interpret what motions it may make afterwards. Without such orientation, subjects have difficulty perceiving the initial movements the face may make. The last frame of each clip was also copied and expanded as a still picture to give each clip a uniform length of two seconds. After editing, each video clip was saved as a Quicktime .MOV file, and its audio portion was copied into an independent .AIFF file for use in the audio-only half of the experiment.

Both video and audio clips were presented to subjects via a computer monitor and headphones in a sound-proof booth. The experiment's twenty-eight subjects were split evenly into audio-visual and audio-only groups. In the video half of the experiment, subjects would see video on the computer monitor while the corresponding audio played over the headphones. In the audio-only condition, the computer monitor went blank while the audio played over the headphones. After the subject had listened to each clip, the computer presented them with the following question: "What word did you hear?" and the subject would respond by clicking on one of three VC or CV alternatives (written on the screen in realistic English spellings), which differed only in the place of articulation of their stop consonant. After the subject had made a selection, they were given the option of either changing their selection or moving on to the next stimulus. Listeners heard the next token only after they had decided to move ahead with the experiment.

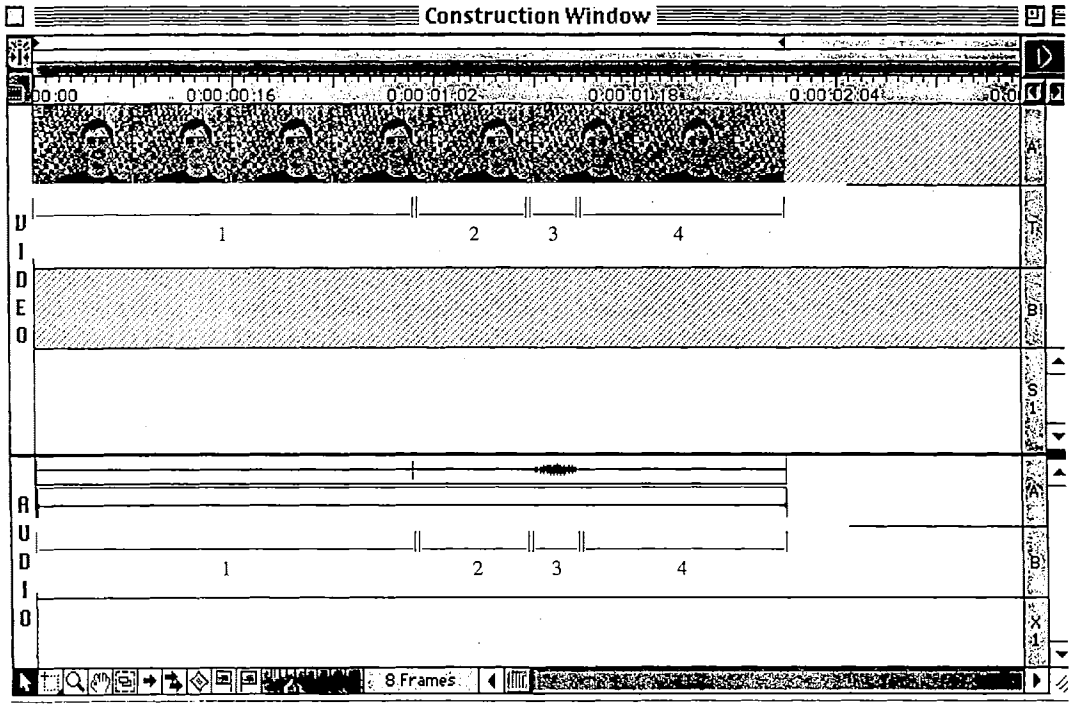

Figure 1: Four part video editing construction of "da" stimulus

1. One second still shot
2. Ten frames prior to onset of acoustics
3. Acoustic production of "da"
4. Expanded still shot to lengthen entire video to two seconds

The tokens were split up into groups with uniform manner, syllabic position and vowel features. This was done so that the listeners would only have to make a decision about the perceived place of articulation of any given token. Within each block, then, the tokens were evenly split between male and female productions, stressed and unstressed productions, and labial, coronal and dorsal productions. The listeners also heard each token twice, so they heard a total of twenty-four tokens in every block. The blocks were evenly split between nasal and oral stops, onset and coda position, and productions with the vowel /i/ or the vowel /a/. This amounted to eight blocks in all, which meant that each experimental trial required the subject to make a place categorization for 192 different tokens. In order to gauge the effects of adding audio information to the speech signal, subjects first worked through all 192 different tokens at their speech reception threshold, and then later repeated the same experiment with the volume at a comfortable listening level.

A person's "speech reception threshold" is the volume level at which that person can understand one-half of the spondees that they hear. In the first stage of this experiment the speech reception threshold of each listener was determined with an adaptation of the method of Cutler and Butterfield (1992). Listeners were isolated in a sound booth, under exactly the same conditions in which they would be presented with the audio-visual stimuli in the second half of the experiment. In the sound booth they listened to a series of six spondees over a set of headphones. After each spondee, the listeners were prompted by the computer to type in what word they thought they heard; after they had responded, they would hear another spondee, and so on. After six spondaic tokens, the number of correct responses would automatically be tabulated by the computer and shown to both the listener and the experimenter. Initially, listeners were familiarized with this task with the volume on the headphones at a comfortable listening level; after their first run, however, the volume would be significantly decreased to a level at which pilot testing had shown most people begin misunderstanding words. After this second trial, the volume was increased or decreased accordingly until the listeners responded correctly to 3 ± 1 of the 6 words they had heard. At this 50% comprehension level, the volume was considered to be at the listener's speech reception threshold.

After this pre-test had established the listener's speech reception threshold, the listener began working through the blocks of stimuli at this volume level. After the

listener had worked through all 192 stimuli, the volume was returned to the original comfortable listening level and the listener repeated the experiment again. There were fourteen subjects in each condition; the subjects were volunteer students from introductory psycholinguistics and linguistics classes. Most were remunerated for their participation and the rest participated for extra credit in their respective courses. All participants were encouraged to take breaks whenever they felt they needed one.

RESULTS AND ANALYSIS

The task that listeners were asked to perform in this experiment was simple: identify one word out of three alternatives as the word they had heard or seen being spoken. Adding a stronger or more salient cue for a certain place of articulation should have two effects with respect to this task--first, it should increase the likelihood that listeners will respond appropriately when they perceive that cue, and secondly, it should decrease the likelihood that listeners will respond incorrectly when they do not hear that cue. When a listener does respond correctly in this task, he or she has, in the battleship-like terminology of speech perception research, scored a "hit". On the contrary, when they mistakenly respond with one alternative when the stimulus was intended as another, they have registered a "false alarm" in some phonological firehouse in their minds. With stronger cues, then, their probability of registering "hits" should increase while their probability of registering false alarms should decrease. Mathematically speaking, this amounts to

$$(11) \quad I = P(hit) - P(fa),$$

where I is a measure of listener "sensitivity"--i.e., how much of an impression an external stimulus makes on a listener. Adding one to this equation and dividing the entire sum by two yields a variable that ranges from 0 to 1:

$$(12) \quad I = \frac{1 + P(hit) - P(fa)}{2}$$

Green and Swets (1966) derived this definition of I in developing a non-parametric equivalent of d' in their "Signal Detection Theory". Calculating listener sensitivity in this way--instead of simply measuring hit rates--helps eliminate listener bias effects by taking the probability of false alarms into account.

Equation (12) was used to calculate sensitivity values for every token by subject in an attempt to quantitatively determine which place cues were more salient than others. Across all conditions sensitivity to the labial place of articulation was highest. Ultimately, labials came in with a sensitivity ranking of .9, followed by dorsals with .83 and coronals at .81 (see Figure 2). A repeated measures ANOVA showed that the place factor was significant (see Appendix 1, #10).

Interestingly, the perceptual strength of labials is not simply an artifact of their strong visual cues. Breaking down sensitivity values for both the audio-only and audio-visual groups of listeners, labials still came out on top in both conditions (Figure 3; #11 in Appendix). With audio-only stimuli, labials are still slightly (but not significantly) higher than dorsals, and in audio-visual stimuli, the labials' sensitivity ranking approaches ceiling while coronals and dorsals are essentially even. The story remains the same once the results are broken down by volume level (Figure 4; #3 in Appendix). At both speech reception threshold and comfortable listening level, labials again show the highest sensitivity, followed by dorsals and coronals.

These findings seem to contradict the previous suppositions of Jun, who claimed that dorsal stops ought to have more salient cues than labials because of their characteristic velar pinch in the transition from articulatory closure to full vocalic opening. It also causes problems for Hume's claim that labials were "perceptibly vulnerable" because of their lack of a salient release burst (as was hypothesized by Ohala in earlier work). These results seem to show that, on the contrary, labials have the most salient cues of any stop place of articulation.

The same results seem less problematic, though, when only the audio group is taken into account (as in Figure 4b). Here labials—without the strength of their visual cues--only have a slight advantage over dorsals in the comfortable listening level condition, and no significant difference exists between them at speech reception threshold. Coronal sensitivity, on the other hand, sinks lower than both dorsals and

labials. These results seem more in line with Jun's original rankings of place cue salience, even though labials are still surprisingly strong.

But comparing how results *change* between audio-visual conditions is the best way to determine which cues (for which place) really contribute the most to listener sensitivity (and could therefore be considered the most *salient* cues). In breaking the results down in this way, it is possible to see why phonologists like Jun and Hume might have made the assumptions they did. Figure 5 shows how much salience *increases* for each place of articulation whenever visual or audio information is added to the signal. These results are interesting for a number of reasons; first of all, the strength of visual cues for labial stops is dramatic, increasing salience values by .18 on the whole. Not quite as dramatic but no less significant is the fact that the salience of coronal stops increases much more than the salience of dorsals does. (.12 vs. .07) Even though coronal stops do not usually induce a McGurk effect, it seems that people are more sensitive to their visual cues than they are to dorsal visual cues.

Adding audio information to the perceptual task seems to turn things around completely, interestingly enough. Dorsals and coronals both gain significantly more salience from the addition of audio information than labials do. Since phonologists have traditionally thought of perceptual salience as limited to a speech event's acoustics, this graph may explain why labials have always gotten the short shrift in past evaluations of perceptual salience. Even though labials are, in general, more salient than coronals or dorsals, they do not seem to gain much salience through only their acoustic cues. If these were the only cues that mattered in the perception of stops, then labials might, indeed, be the most "perceptibly vulnerable" of the various places of articulation. It may also be the case that only acoustic cues have an "outside-in" effect on phonological structure.

Part of what might have reduced labial sensitivity in these comparisons, though, is the ceiling effect induced by the comparative strength of the labials' visual cues. Since labials in the audio-visual condition approximate maximum sensitivity, there is little room left for them to improve when more audio information is added to the speech reception threshold condition. Figure 5b shows a slightly modified version of Figure 5, calculating the increase in audio sensitivity by only including the differences between the two audio-only conditions. Here added audio information increases the sensitivity of labials just as much as it increases the sensitivity of coronals or dorsals. This figure

probably paints a more realistic picture, therefore, of the audio-only strength of labial cues in comparison to other places of articulation. Since there are no significant differences among acoustic cues for the three places of articulation, it seems difficult to claim that they might drive phonological rules applying to one place of articulation but not the others.

Almost all of the other factors tested in this experiment yielded significant results that might have been predicted by those familiar with phonological theory and speech perception. Besides the between-subjects video factor and the within-subjects volume factor, the syllabic position of the stop consonant also contributed significantly to cue salience. Stops in onset position were more salient than stops in coda position, in other words (Figure 6; #6 in Appendix). These results confirm Jun's ranking of preservation constraints for coda and onset position in (4). Another significant factor was stress, which implies that place cues were more salient in stressed syllables than in unstressed syllables (Figure 7; #7 in Appendix). This confirms Hume's conjecture that place cues are more salient in stressed syllables than unstressed syllables; this may, therefore, be one motivating factor in metathesis processes (as in Kui) that shift labials from unstressed to stressed syllables. However, labials lose salience in moving from onset to coda position (see Figure 6), so perceptual gain is probably not a factor in metathesizing labials between these positions.

Interestingly, the one factor which did not prove to be significant was the manner factor--sensitivity did not significantly increase in oral stops as opposed to nasal stops (Figure 8). Although sensitivity did increase somewhat between these two conditions, its F value fell just short of reaching the 1% significance level in the repeated measures ANOVA (df=1,27, F=3.375, p=.078). This result is surprising in that it contradicts Jun's ranking in (5), in which he claimed that cues for oral stop place are stronger than cues for nasal stop place. It also seems surprising given the relative susceptibility of nasals to undergo place assimilation (see Mohanon (1993)).

Figure 8 also provides some explanation for the strong position*manner factor (df=1,26, F=215.032, p=.000; #16 in Appendix). Figure 8 shows that, even though manner alone was not a significant factor, there *was* a significant difference between the sensitivity of oral dorsal stops vs. nasal dorsal stops. This difference probably arises from the fact that half of the dorsal nasal stops in this experiment were in onset position,

which is not allowed by English phonotactics. The English-speaking listeners in this experiment were therefore forced to make perceptual judgments about dorsal nasals in a completely unexpected syllabic position; their failure to perceive these segments as well as they perceived their oral counterparts may be attributed to their lack of experience in dealing with such a perceptual task. This discrepancy also reveals the insidious persistence of inside-out phonological effects even in this minimally meaningful experimental task. The fact that phonological knowledge contributed to listeners' perception of phonotactically acceptable sequences means that the judgments the listeners made in this experiment were not simply universal responses to the *inherent* cues for the different places of articulation. A true evaluation of the strength of these inherent cues would have to find some way to eliminate these language-specific phonological effects.

Neither Jun nor Hume mentioned vowel-specific effects on patterns of phonological assimilation or consonant/consonant metathesis, but this study included consonant productions with both /a/ and /i/ on a hunch that visual effects might be stronger for a more open vowel (like /a/) than for a more closed vowel (like /i/). There was a significant vowel effect in the repeated measures ANOVA (df=1,26, F=135.744, p=.000; #4 in Appendix), but this apparently had more to do with the acoustic characteristics of /a/ and /i/ than it did with their visible effects on consonant articulation. /a/ had a much higher inherent amplitude than /i/, and therefore induced much higher sensitivity scores in the audio-only conditions. In the audio-visual conditions, however, these acoustic effects disappeared and productions with /a/ and /i/ were perceived equally well. The vowel*video factor is thus significant (df=1,26, F=156.887, p=.000; #5 in Appendix), but for reasons that were not originally expected.

DISCUSSION

One reason that this experiment yielded such surprising results--and failed to justify cross-linguistic patterns in metathesis and place assimilation--may be that it oversimplified the experimental task. Though Jun and Hume both refer to the "inherent" perceptual salience of segments in motivating their phonological hypotheses, they are both concerned with processes that take place in a particular phonological environment. Hume, for instance, is concerned with consonant/consonant metathesis across a syllable

boundary while Jun is mostly preoccupied with stop place assimilation in coda position. In an effort to simplify the perceptual task (and also eliminate potential "inside-out" influences on perception), this experiment only tested the perception of place in an isolated context in nonsense words. It did not strictly test unreleased stops or stops that were immediately followed by conflicting place information for some other consonant. It is very likely that the relative salience of certain stop cues may change in these different contexts, and it may be that this variation in salience is what motivates certain assimilatory and metathesis processes in a language's phonology. Testing this contextual salience in such a way that listeners cannot depend on internalized language-specific knowledge about place cues in context but must, rather, base their perceptual judgments only on what sounds they hear or see seems to be a daunting task for speech perception. However, only with such studies could the universal facts about place cue perception (in or out of a linguistic context) be established and thereafter used with any scientific certainty in phonological analyses.

On a more immediate note, this present study offers a new insight into the inherent salience of audio and visual cues for stop place of articulation. Some of its most interesting results involve the strength of both audio and visual cues for labial stops. The perceptual significance of visual cues for coronal stops also seems to contribute something new to our knowledge of visual speech perception, since these cues do not seem to be strong enough to induce a "McGurk effect" and have therefore gone hitherto unrecognized. The work of Hume et al. (1999) also shows that the salience of acoustic dorsal cues increases greatly when they are produced with the vowel /u/, which was not included in this study. A future replication of this study with more and different vowels may give reason to re-evaluate the tentative ranking of cue salience by place.

Hume et al. (1999) also shows that speakers of different languages may vary in sensitivity to different acoustic cues for place. Likewise, some studies by Sekiyama and Tohkura (1991 and 1993) show that the strength of the McGurk effect may differ between Japanese and American listeners. The fact that such cross-linguistic differences in perception seem to exist makes it impossible to claim that the English-only results of this experiment genuinely reflect some universal tendencies in perception. Replicating this experiment with native perceivers of other languages is only one of the many tasks that

need to be undertaken by those theorists who deem it necessary to turn phonology inside out.

References

Cutler, Anne and Sally Butterfield (1992) Rhythmic cues to speech segmentation: evidence from juncture misperception. *Journal of Memory and Language*, **31**, 218-236.

Green, D.M. and Swets, J.A. (1966) *Signal Detection Theory and Psychophysics*. New York: Wiley.

Hume, Elizabeth (1998) The role of perceptibility in consonant/consonant metathesis. To appear in: *WCCFL XVII Proceedings*. Eds. Susan Blake, Kim Duk-Sook, and Kimary Shahin. Stanford: CSLI.

Hume, Elizabeth, Keith Johnson, Misun Seo, and Georgios Tserdanelis (1999) A cross-linguistic study of stop place perception. To appear in: *Proceedings of the 14th International Congress of Phonetic Sciences*.

Jun, Jongho (1995) Place assimilation as the result of conflicting perceptual and articulatory constraints. *WCCFL XIV Proceedings*, 221-237.

Kohler, K.J. (1990) Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In W. J. Hardcastle and Alain Marchal (eds.), *Speech Production and Speech Modelling*. 69-92. Dordrecht: Kluwer Academic Publishers.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. *Language & Speech*, **6**, 172-87.

Malécot, André (1958) The role of releases in the identification of released final stops. *Language*, **34**, 370-380.

McGurk, Harry and John W. MacDonald (1976) Hearing lips and seeing voices. *Nature*, **264**, 746-748.

Miller, G.A. and Nicely, P.E. (1955) Analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338-52.

Munhall, K.G., Gribble P., Sacco L., Ward M. (1996) Temporal constraints on the McGurk effect. *Perception & Psychophysics* **58 (3)**, 351-362.

Mohanon, K.P. (1993) Fields of attraction in phonology. In John Goldsmith (ed.), *The Last Phonological Rule*. 61-116. Chicago: The University of Chicago Press.

Ohala, John (1996) Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, **99 (3)**, 1718-1725.

Schaengold, Charlotte (1999) English and Navajo Languages in Contact. Unpublished manuscript, The Ohio State University.

Sekiyama, Kaoru and Yoh'ichi Tohkura (1991) McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, **90**, 1797-1805.

Sekiyama, Kaoru and Yoh'ichi Tohkura (1993) Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, **21**, 427-444.

Stevens, Kenneth (1989) On the quantal nature of speech. *Journal of Phonetics*, **17**, 3-45.

Strand, Elizabeth and Keith Johnson (1996) Gradient and visual speaker normalization in the perception of fricatives. In Dafyold Gibbon (ed.), *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference*, Bielefeld, October 1996. 14-26. Berlin: Mouton de Gruyter.

Wang, M.D. and Bilger, R.C. (1973) Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America*, **54**, 1248-66.

Warren, R.M. (1970) Perceptual restoration of missing speech sounds. *Science*, **167**, 392-393.

Table I
Significant effects from a repeated measures ANOVA of sensitivity (I)

Source of Variance	DF	F	P	
Position*Manner	1,26	215.032	0.000	
Vowel*Video	1,26	156.887	0.000	
Vowel	1,26	135.744	0.000	
Place	2,25	115.32	0.000	(see Figure 2)
Volume	1,26	104.204	0.000	
Vowel*Manner	1,26	34.534	0.000	
Manner*Stress	1,26	34.046	0.000	
Vowel*Position*Video	1,26	31.098	0.000	
Position*Manner*Place*Video	2,25	28.673	0.000	
Vowel*Place*Video	2,25	27.613	0.000	
Place*Video	2,25	26.586	0.000	(see Figure 3)
Position*Manner*Place	2,25	26.586	0.000	
Position*Stress	1,26	25.271	0.000	
Volume*Video	1,26	23.522	0.000	
Volume*Position*Manner	1,26	22.477	0.000	
Volume*Vowel*Place*Video	2,25	17.799	0.000	
Volume*Position*Place*Video	2,25	17.073	0.000	
Vowel*Place	2,25	16.196	0.000	(see Figure 4)
Volume*Vowel*Place	2,25	13.795	0.000	
Volume*Position*Place	2,25	12.238	0.000	
Stress	1,26	15.382	0.001	(see Figure 7)
Vowel*Position	1,26	15.078	0.001	
Volume*Place	2,25	9.023	0.001	
Vowel*Manner*Video	1,26	12.342	0.002	
Manner*Place	2,25	7.878	0.002	
Vowel*Manner*Stress*Place	2,25	7.715	0.002	
Position	1,26	10.190	0.004	(see Figure 6)
Vowel*Position*Manner	1,26	9.266	0.005	
Volume*Manner*Stress*Video	1,26	7.779	0.010	
Volume*Place*Video	2,25	5.519	0.010	(see Figure 5)

Between listeners factor:
Video: Audio-visual, Audio-only

Within listeners factors:
Place: Labial, Coronal, Dorsal
Volume: Speech reception threshold, Comfortable listening level
Position: Onset, Coda
Stress: Stressed, Unstressed
Manner: Oral stops, Nasal stops
Vowel: [a], [i]

Figure 4b: SRT vs. CLL Sensitivity
(Audio Group Only)

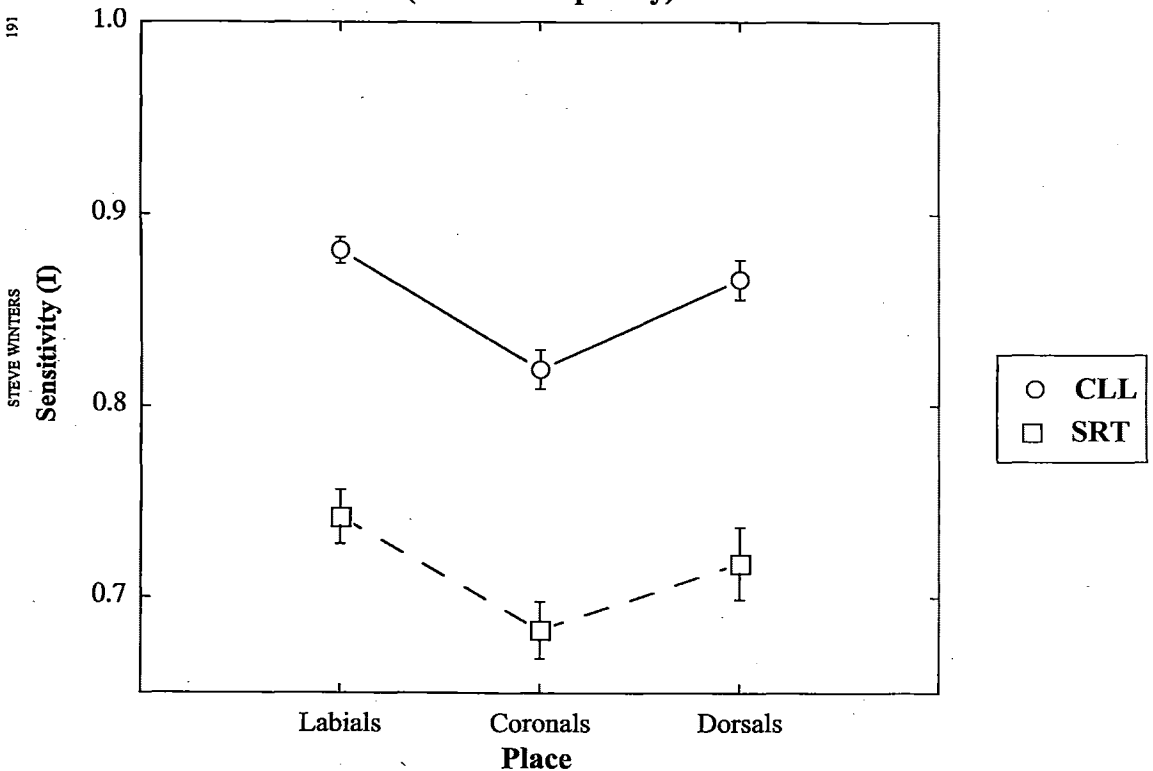

Figure 5: Audio and Video Contribution to Sensitivity

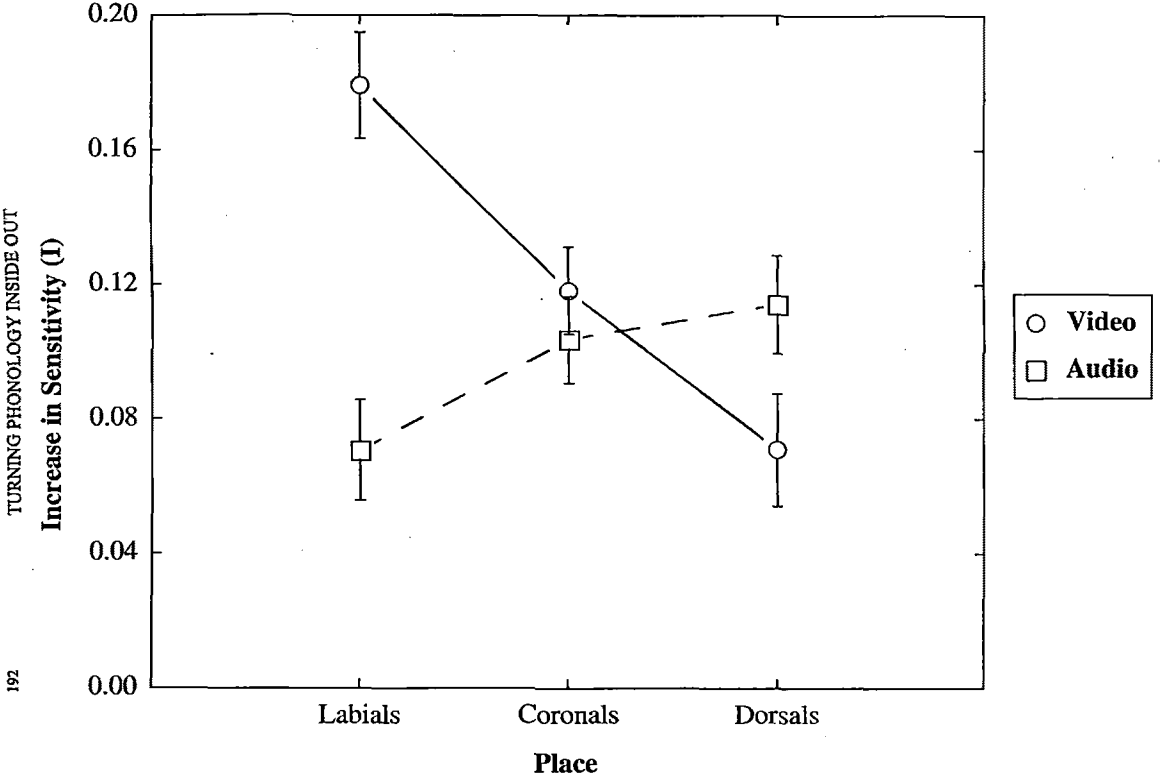

Figure 5b: Audio (only) and Video Contribution to Sensitivity

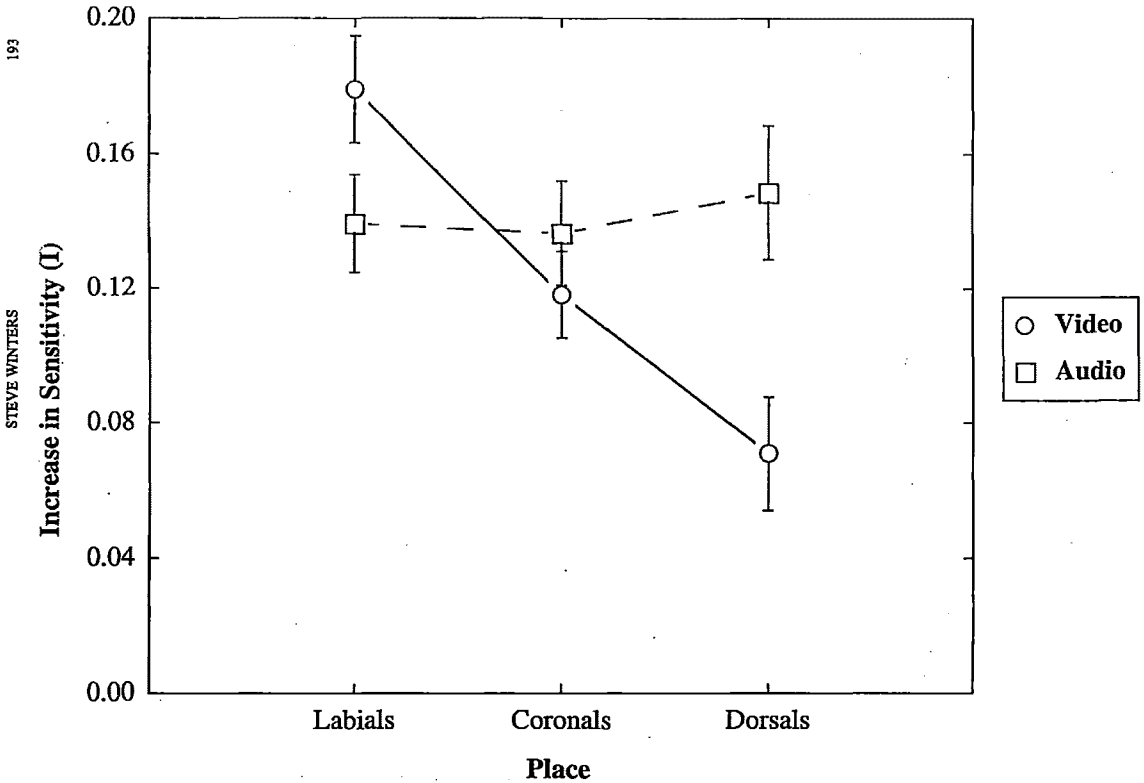

Figure 6: Onset and Coda Sensitivity

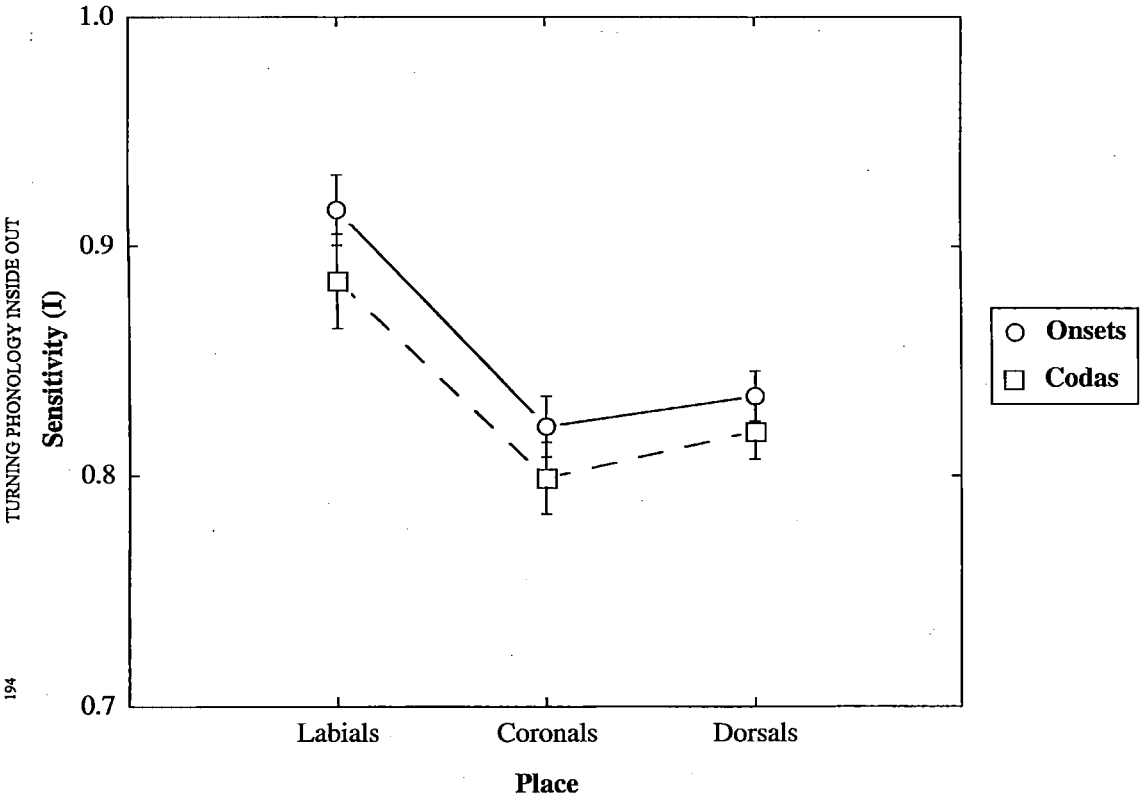

Figure 7: Stressed vs. Unstressed Sensitivity

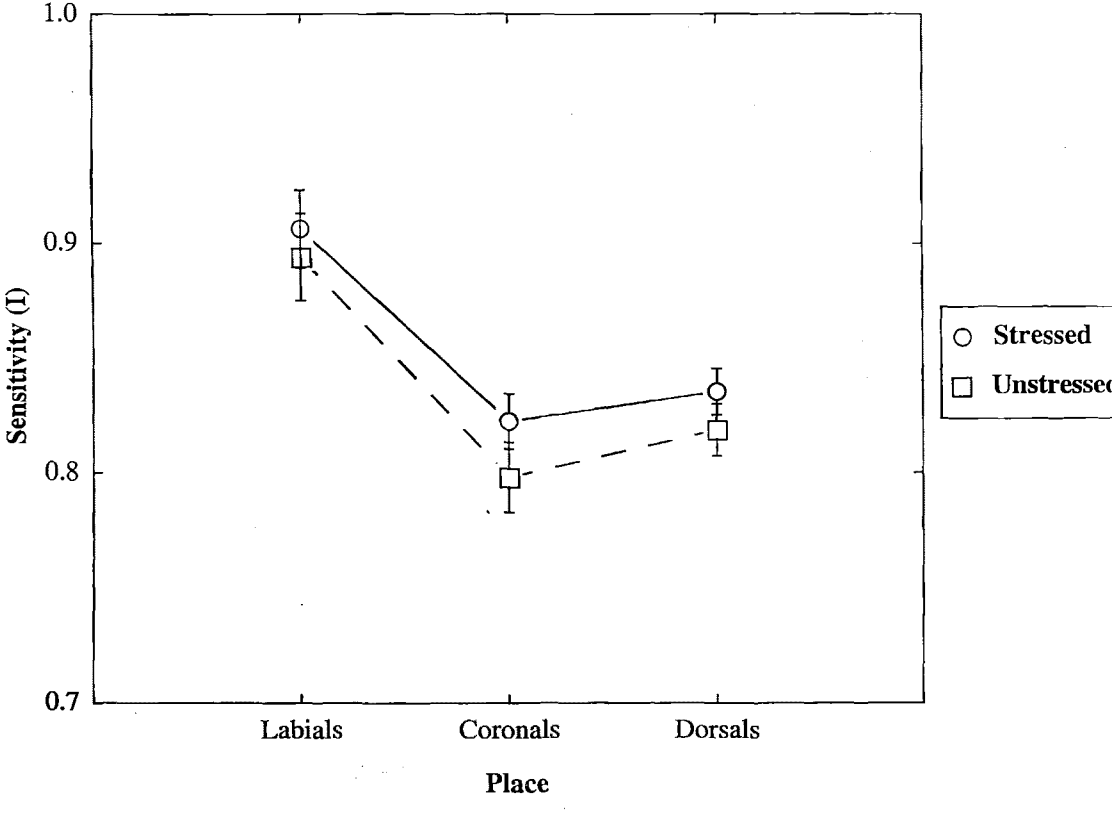

Figure 8: Stops vs. Nasals Sensitivity

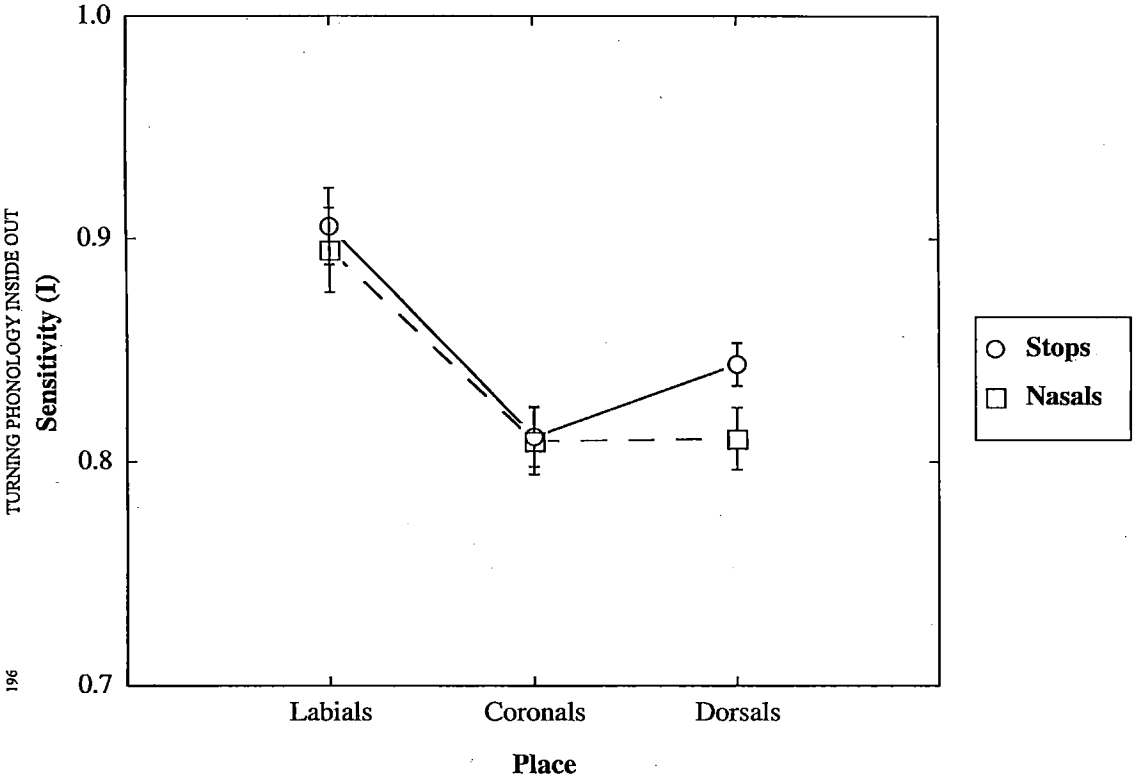

Figure 2: Sensitivity across all conditions

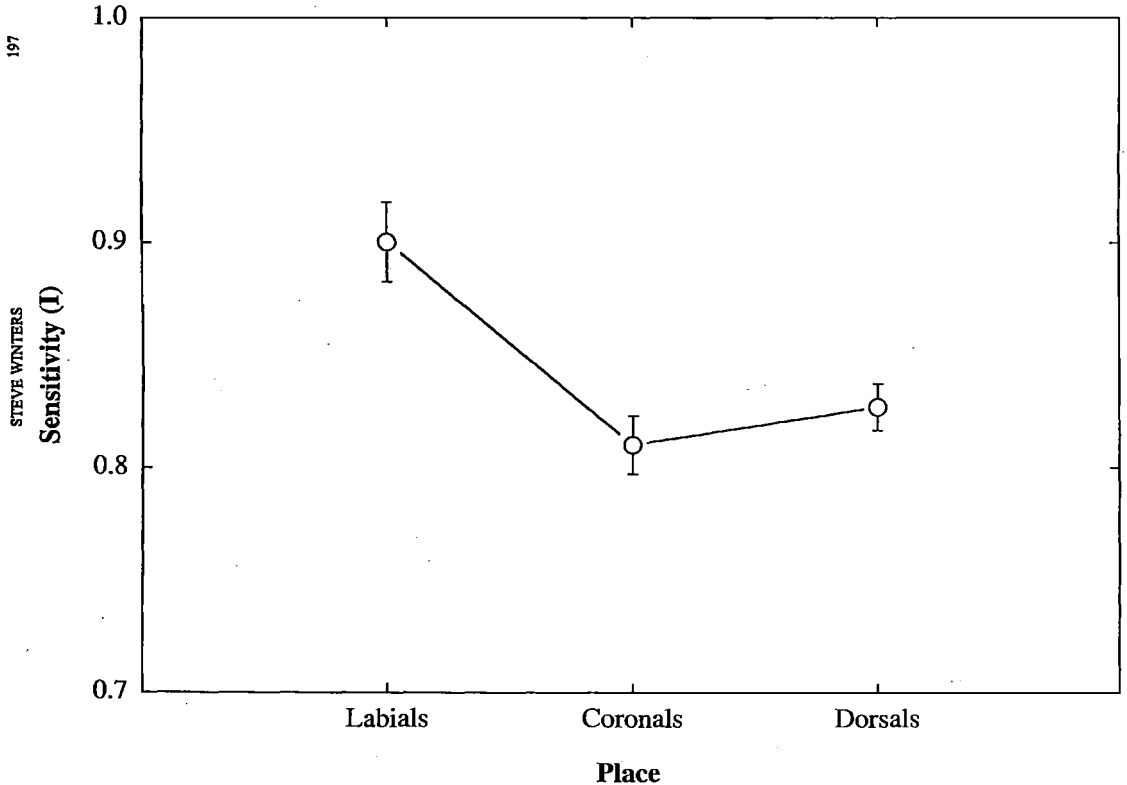

Figure 3: Video vs. Audio Sensitivity