OSU #36: Papers from the Linguistics Laboratory 1985-1987, Sept. 1987.

Perceiving by syllables or by segments: Evidence from the perception of subcategorical mismatches

Keith Johnson
Linguistics, Ohio State University

Abstract: This paper describes an experiment in which two general hypotheses concerning speech perception are tested. According to the segment perception hypothesis the acoustic signal is interpreted in terms of segments analogous to those used by phoneticians in transcribing speech. The syllable perception hypothesis on the other hand holds that the speech signal is perceived in terms of syllable sized units. The experiment tests these two hypotheses by presenting subjects with a perceptual task for which the two make opposite predictions. Tokens with subcategorical mismatches were produced by cutting the fricatives [s] and [ʃ] from VC syllables (vowels were [i,a,o,u]) and recombining them with vowels which differed from the original context in terms of transitions and rounding. The segment perception hypothesis predicts that in syllables with transition mismatches (ie. transitions for [s] and with [ʃ] actually occurring) coarticulatory rounding on the actually occurring fricative will aid in the perception of [ʃ] and slow the perception of [s], while the lack of rounding on the actually occurring fricative will have the opposite effect. This is because the rounding makes [ʃ] a more extreme example of [ʃ] (and thus easier to categorize as such) while rounding makes an [s] less distinctly an [s]. The syllable perception hypothesis predicts that in syllables with transition mismatches coarticulatory rounding on the actually occurring fricative will aid the perception of [s] and hinder [ʃ] perception. This is because the [s] with rounding is acoustically closer to the prediction made on the basis of the transition on the vowel. Similarly, the [ʃ] with rounding is acoustically further removed from the [s] which is expected as a result of the transitions on the vowel in a mismatched syllable and thus should require more time to be perceived as [ʃ]. The results of the experiment reported here support the segment perception hypothesis. Subjects' perception of [s] in syllables with transition mismatches was inhibited by coarticulatory rounding while their perception of [ʃ] in syllables with transition mismatches was facilitated by coarticulatory rounding.

1. Introduction

The experiment described in this paper was designed to test two hypotheses about speech perception. These two hypotheses will be called syllable perception and segment perception. As their names indicate they differ in so far as they entail that the basic units of speech perception are respectively syllables and segments. Advocates of the first approach

include Klatt (1980)[1], Massaro (1972), and Morton and Broadbent (1967). The theorists who suppose that segments are perceived include Bondarko et al. (1970), Fant (1967), Stevens and Halle (1967), Liberman et al. (1967), and Pisoni and Sawusch (1975). As these lists indicate there are a wide variety of ways that syllable or segment perception might be conceived.

The unifying feature of the different perception by segments approaches is that they all hold that the objects of speech perception are phonemes and that subsequent percepts (syllables or words) are sequences of phonemes. Figure 1 shows the sequence of perceptual events as envisioned in a perception by segments approach.

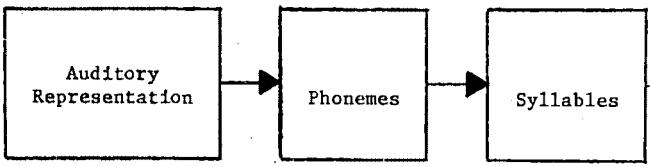

Figure 1: Organization of the speech perception process from a perception by segments approach.

In this kind of model the recognition device takes as input a preliminary auditory analysis and computes as output phones. Pisoni and Sawusch (1975) proposed such a model of perception. In their view speech perception is accomplished via (1) acoustic feature analysis, (2) phonetic feature analysis, (3) a feature buffer, and (4) phonetic feature combination. The key element in the segment perception approach is that perception is accomplished in terms of units which correspond to the symbols a phonetician might use to transcribe the utterance. Thus, in the segmental model the things being perceived are segments which are then combined into syllables.

An example of the syllable perception approach is found in Klatt (1980) and is illustrated in figure 2. This figure shows a phonetic decoding network. The network defines possible sequences of spectra. When the perceptual system matches a particular sequence (i.e. a particular path through the network is followed) the syllable defined by the sequence is perceived. There is no intermediate perceptual stage between auditory analysis and identification of a syllable. In this model the identification of component phonemes can only be accomplished after the entire syllable is identified.

[1]Klatt's (1980) model actually includes both a version of syllable perception and segment perception. Since he argues in the body of his paper for syllable perception I am including him in that camp.

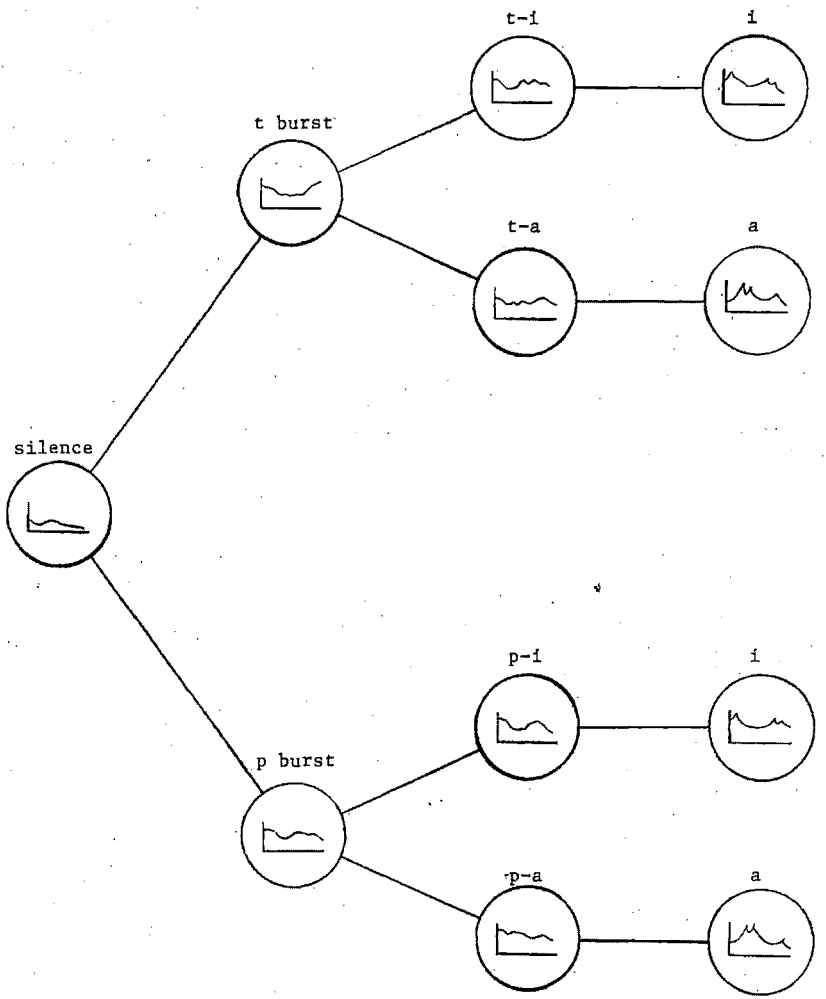

Figure 2: A spectral decoding network for the syllables [ti], [ta], [pi], and [pa].

The experiment reported here used naturally produced speech tokens which were edited to produce subcategorical mismatches, following the technique used by Whalen (1984). The subcategorical mismatches used in this experiment were produced by cutting VC syllables and recombining the resulting segments with V's or C's from other syllables. More specifically, the fricatives [s] and [ʃ] were produced after four different vowels [i,a,o,u]. The mismatches which resulted from recombining these segments were of two types. Transition mismatches resulted when vowels from the context [__s] were recombined with [ʃ] or when vowels from the context [___] were paired with [s]. Rounding mismatches resulted when fricatives from the context [V_{+rnd}__] were recombined with [-rnd] vowels, or when fricatives from the context [V_{-rnd}__] were recombined with [+rnd] vowels.

Mismatches pose an interesting problem for speech perception theories because coarticulatory information remains in the segments which are separated from each other. In the case of transition mismatches the discrepancy between the place of articulation information in the vowel and that in the fricative itself produces a fairly large and stable reaction time lag in perception (Whalen 1984). It is also the case, though, that the discrepancy between the rounding of the vowel and the effects of rounding coarticulation in the fricative could affect perception.

When subjects are asked to identify the fricative noise in mismatched tokens such as these it is often the case that they respond before the completion of the acoustic signal. Due to the fact that the subject's reaction time involves both perception time and response time it is very likely that the subject has established some predictions concerning the identity of the fricative during the vowel - based on the transitions and the roundness of the vowel. Predictions such as these seem also to be the most plausible explanation of the effect of subcategorical mismatches on reaction time in identifying the fricative in VC syllables. The vowel portion of the syllable allows the listener to set up some expectations concerning the following fricative. When the expectations are not met identification is slowed.

The two hypotheses characterize the listener's predictions in quite different ways. In the syllable perception model it must be assumed that the hearer makes predictions which are below the level of segmental identities. The predictions are acoustical in nature because within a syllable the perceptual system is seen as progressing from one spectral template to the next. In the segmental model predictions are made in terms of categories instead of acoustic values. For instance, the occurrence of [u] leads to a prediction that the following consonant will have rounding.

Figure 3 illustrates the syllable model for the experimental tokens. This figure is analogous to figure 2. If we suppose that the vowel [u] from [uʃ] is presented, then the prediction for the following state in the network is the spectral template for [ʃ]. If the speech token being presented is a mismatched token which has [s] instead of [ʃ] then the fact that the perceptual prediction was a spectrum suggests a strategy for recovering from the mismatch. The general requirement is that another spectral template be found which will match the actual fricative spectrum.

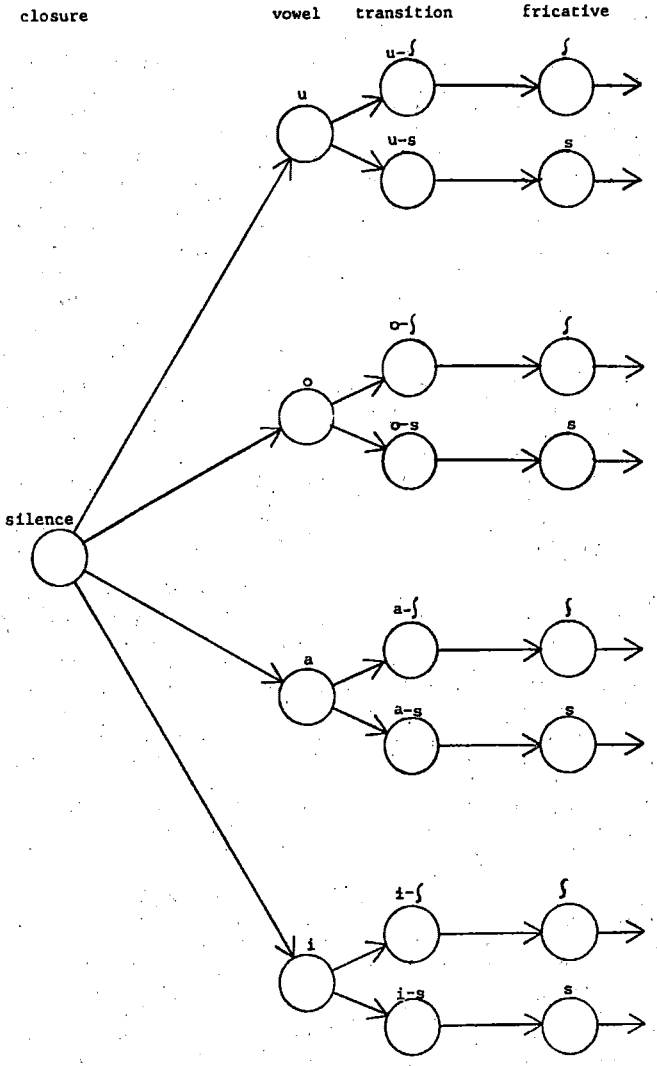

Figure 3: Perception of the VC test tokens in a transition network.

One way for the syllable perception model to recover from a mismatch is for the perceptual mechanism to attempt to interpret the auditory presentation as a well-formed syllable by revising the template or the auditory representation of the sound or both. Thus, if the expected fricative is [ʃ], rounded [s]'s will be more easily perceived because their spectrums are closer to the expected spectrum (i.e. they involve less revision of templates and/or representations). This method of recovery would predict that if the hearer sets up an expectation for an [s]

spectrum, then it will be easier to recover from a mismatch involving the [ʃ] of [iʃ] than one involving the [ʃ] of [uʃ] because the spectrum of the [ʃ] of [iʃ] has more in common with an [s] spectrum than does the spectrum of the [ʃ] of [uʃ] (see figure 3). If the hearer sets up an expectation for an [ ] spectrum the reverse is true. The [s] spectrum from [us] would be easier to process while the [s] spectrum from [is] would be more difficult.

The segment perception hypothesis also leads to some predictions about how the hearer might recover from a mismatch. If segment information which is spread out over the syllable is integrated in the process of forming a segment identification, then the conflict between cues in the mismatched cases will have to be resolved. One way that a resolution between conflicting cues might be reached is by comparing the relative strengths of the cues. This method of recovery also results in predictions for the relative ease of processing the fricative mismatches in this experiment. Regardless of the transition cues the fricative that will be easiest to process as an [s] is the [s] of [is]. This is because this sound is the most extremely s-like [s] of the set. When the relative strengths of the cues for the final consonant are compared this 'strong' [s] will over-ride the misleading information in the transition more quickly than will the [s] from [us]. This same type of situation prevails when a vowel with alveolar transitions is paired with an alveopalatal fricative. The [ʃ] from [uʃ] will be easier to process because it is less like an [s] than any of the other [ʃ]'s.

Thus, the two theories make opposite predictions about the ways in which rounding in the fricative will help or hinder the preception of transition mismatched tokens. The syllable perception model predicts that coarticulatory rounding will make [s] easier to perceive when [ʃ] is expected, while the segment perception approach predicts that rounding will hinder the perception of [s] when [ʃ] is expected. The effect of rounding in [ʃ] perception has the opposite pattern of predictions. In perceiving by syllables, rounding an [ʃ] should inhibit reaction time while the perceiving by segments approach predicts that rounding should facilitate reaction time.

2. Methods

The tokens used were constructed from the syllables [us, os, as, is, uʃ, oʃ, aʃ, iʃ]. These syllables were recorded in an anechoic chamber using high quality equipment. The speaker was a male native speaker of American English. They were then digitally rerecorded at a sampling rate of 15 kHz (low pass filtered at 7 kHz). The digitized forms were edited so that the vowel and fricative portions were separated. The cut was made at the point at which the periodicity of the vowel ceased. In most cases there was a small amount of frication left on the vowel but this was so low in amplitude that it could not be heard when the vowel portion was played by itself.

Figure 4 shows the spectra of the eight fricative sounds used in this experiment (each graph is the average of 10 consecutive spectra from the first half of the fricative).

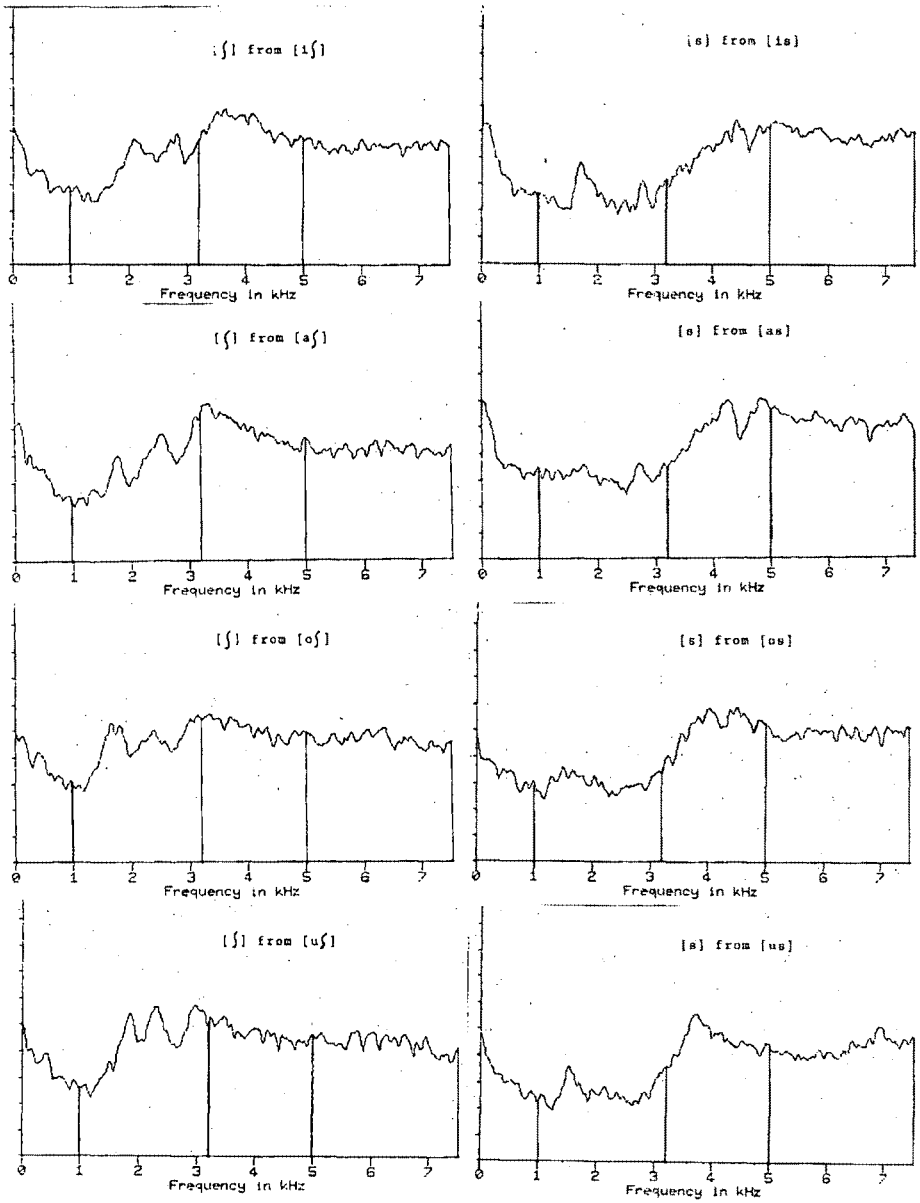

Figure 4: Spectra of the eight fricative sounds.

Using a technique that Jassem (1979) found effective in classifying fricative spectra, the spectra have been broken into regions (1000-3200, 3200-5000, and 5000-7500). Jassem found that by estimating the center of gravity in each of these three spectral regions the fricative can be correctly classified 80-90% of the time. For example, notice the middle region (3200-5000) in the [s] tokens. The center of gravity for the [s] from [is] in this region will obviously be greater than the center point of the region. As lip rounding increases the center of gravity decreases (to less than 4000). It is interesting to note that the center of gravity for the 3200-5000 region is very similar for the [s] from [us] and the [ʃ] from [iʃ]. These two are different in the first region but the similarity in the second is interesting. It makes it possible for us to consider these eight fricatives to be a type of continuum from the [s] of [is] which has a high center of gravity in region 2 and a low center of gravity in region 1, to the [ʃ] from [uʃ] which has a low center of gravity in region 2 and a high one in region 1.

The durations of the vocalic segments of the tokens were comparable to each other (intrinsic vowel length differences were retained). Likewise, the durations of the fricatives and the vowel fundamental frequencies were relatively uniform. This information is in table 1.

Table 1

	V	C	F0
is	226	215	161
iʃ	221	213	159
as	240	201	147
aʃ	238	206	155
os	238	204	157
oʃ	233	215	157
us	214	216	160
uʃ	224	211	156

Durations of the vocalic and consonantal portions of the stimulus items in milliseconds. F0 is in Hz.

Eight vowel tokens and eight fricative tokens resulted from cutting the VC syllables. In order to create the tokens which were used in the experiment each vowel token was combined with each fricative token. This is illustrated in table 2.

Table 2
The 64 tokens used in the experiment.

		vowels							
		u-s	u-ʃ	o-s	o-ʃ	a-s	a-ʃ	i-s	i-ʃ
fricatives	u-s	1	9	17	25	33	41	49	57
	u-ʃ	2	10	18	26	34	42	50	58
	o-s	3	11	19	27	35	43	51	59
	o-ʃ	4	12	20	28	36	44	52	60
	a-s	5	13	21	29	37	45	53	61
	a-ʃ	6	14	22	30	38	46	54	62
	i-s	7	15	23	31	39	47	55	63
	i-ʃ	8	16	24	32	40	48	56	64

i-s stands for the [s] from [is],
i-s stands for the [i] of [is], and so on.
Each VC combination is identified by token number (1-64).

By recombining the vowel and fricative portions in this way subcategorical mismatches are created. The two types of mismatches which are created in this particular case are (1) transition mismatches and (2) rounding coarticulation mismatches. For instance, in the first column in table 2 tokens 2, 4, 6 and 8 have transitions for [s] in the vowel but actually end in [ʃ]. Also, in the first column tokens 5-8 (and to an extent 3 and 4) are mismatched in lip rounding. They have a rounded vowel but a fricative which was originally produced with an unrounded vowel. This particular type of mismatch was central to the experiment here.

Fifteen paid subjects participated in this experiment. All subjects were native speakers of English and none reported any hearing loss. Each subject heard and responded to each of the sixty-four tokens described above four times (four blocks of 64 trials). Twenty practice items preceded the actual experimental trials. The experiment was conducted at the Linguistics Laboratory of The Ohio State University using a New England Digital Able 60 computer and the ERS experiment running package.

Subjects were seated in a quiet listening booth wearing Sennheiser HD420 headphones (the volume had been preset to a comfortable listening level). They were seated in front of a Heathkit VT52 computer terminal and responded to each token by pressing either the <s> key (for [s]) or the <h> key (for [ʃ]). Subjects responded to [ʃ] with their right hand and [s] with their left hand. One result of this arrangement was that an effect for handedness showed up in a main effect for fricative ($F(1,14)=8.923$, $p<.01$). The terminal was also used to provide subjects with reaction time feedback. Feedback during practice items included both reaction time and correct answers to the practice trials. The intertrial interval was 2 seconds.

Following Whalen (1984), only correct responses within a prescribed reaction time range (100 to 1000 ms) were included in the data analysis. The design of this experiment was such that only those tokens with transition mismatches were analyzed (this comes to 32*4 observations per subject). The overall error rate then was 11.25%.

3. Results

A three factor repeated measures analysis of variance was performed on the data collected in this experiment. The three factors were: The 'actual vowel' presented to the subject ([u,o,a,i]), the 'rounding' of the fricative presented (classed by the original vocalic context of the fricatives - [u,o,a,i]) and the 'fricative' sound actually presented ([s,ʃ]).

Figure 5 is a plot of mean reaction times as a function of the three factors. There is one graph for each of the four levels of the 'actual vowel' factor. The abscissa of each plot is used for the 'rounding' factor (four levels), and [s] identification is plotted with a dashed line while [ʃ] identification is represented by a solid line.

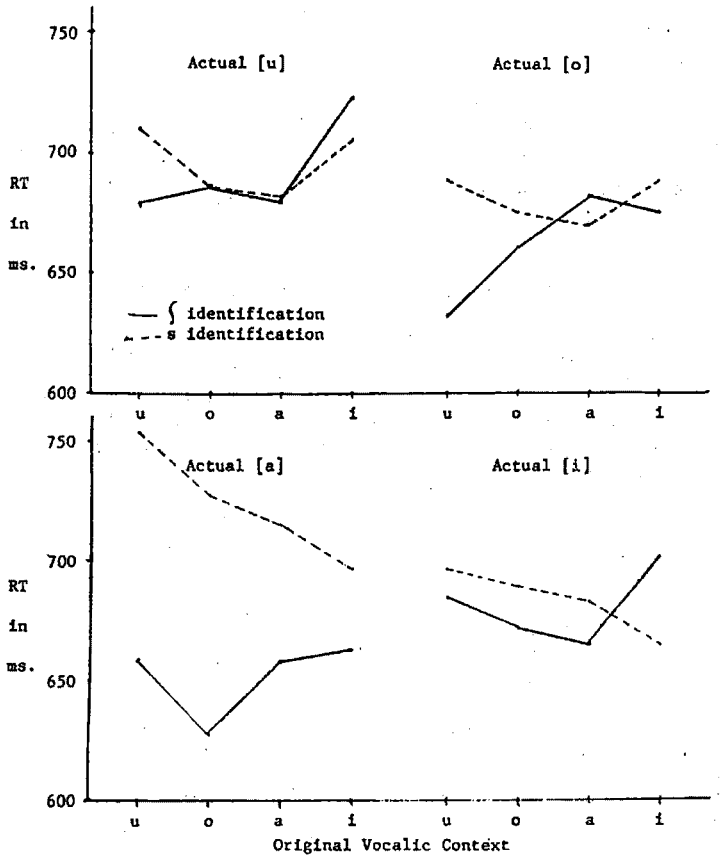

Figure 5: Reaction time to mismatched [s] and [ʃ] by actual vowel and rounding context.

In figure 6 the [ʃ] and [s] identification functions from each of the 'actual vowel' treatments are collected. The [ʃ] identifications tend to be faster when the original context of the fricative was [u], while [s] identification tended to be faster when the original context of the fricative was [i].

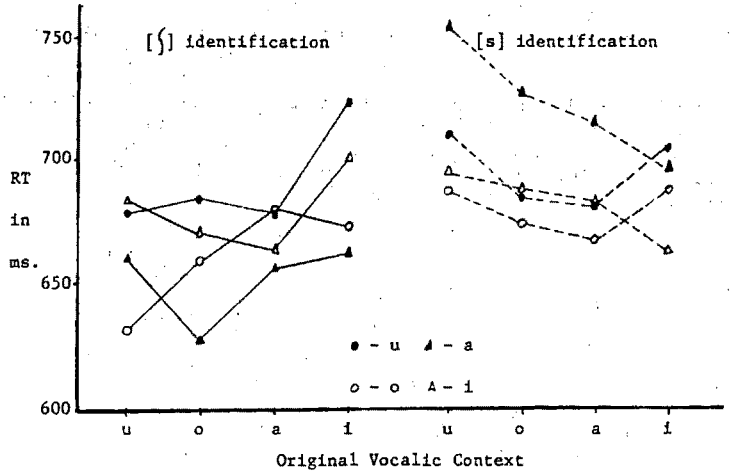

Figure 6: Reaction time to mismatched [s] and [ʃ] grouped by fricative.

When the scores are averaged across the 'actual vowel' factor these tendencies are more easily observable (figure 7). The interaction between the 'rounding' and the 'fricative' conditions (i.e. the functions plotted in figure 7) was significant ($F(3,42)=3.52$, $p < .05$). The direction of this interaction supports the segment perception hypothesis, rather than the syllable perception hypothesis.

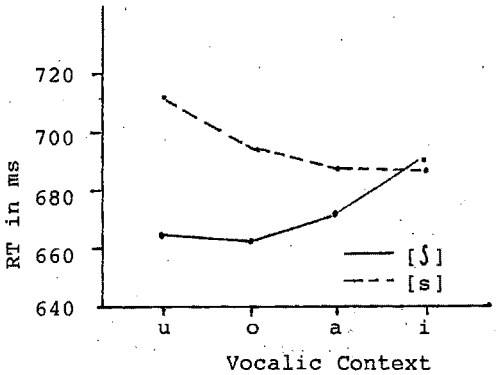

Figure 7: Reaction time to mismatched [s] and [ʃ] by rounding condition.

Two additional conditions proved to have significant results - the main effect for the 'actual vowel' treatment ($F(3,14)=2.8327$, $p<.05$) and the 'fricative' X 'actual vowel' interaction ($F(3,42)=8.169$, $p<.01$). These results are anomalous. None of the hypotheses being tested offer explanations for them. Thus, I will tentatively attribute them to some uncontrolled aspect of the tokens.

4. Conclusion

I've attempted to compare two general classes of speech perception theories. These I called segment perception and syllable perception. It is possible that there are other versions of these hypotheses which would entail different predictions from those tested in this experiment. However, for such alternatives to be useful they must make predictions which are explicit enough to be tested.

The results of this experiment indicate that the class of speech perception theories which entail perception by segments correctly characterize the nature of speech perception (at least in the case of post-vocalic fricatives). This leaves open a wide variety of questions concerning the perception of speech segments. It is still possible to posit active models or passive models, analysis by synthesis or motor approaches. Yet, one thing is suggested by this experiment: syllables in speech are perceived as the result of segment perception, not vice versa.

Acknowledgements

Thanks to Neil Johnson and Rob Fox for their comments and advice at the birth of this project. I appreciate the comments that I've received from Ilse Lehiste, Mary Beckman, Bruno Repp and Dennis Klatt. I'm responsible for my own errors. Thanks also to Debbie Stollenwerk for not talking about linguistics and to Keith Green for musical assistance.

References

Bondarko, L. V.; Zagorujko, N. G.; Kozhevnikov, V. A.; Molchanov, A. P. and Chistovich, L. A. 1970. A model of speech perception in humans. Working Papers in Linguistics. no. 6, Computer and Information Research Center, Ohio State University.

Fant, C. G. M. 1967. Auditory patterns of speech. in Wathen-Dunn, W. (ed) Models for the Perception of Speech and Visual Form. Cambridge, Mass.: MIT Press.

Jassem, W. 1979. Classification of fricative spectra using statistical discriminant functions. in Lindblom, B. and Ohman, S. (eds) Frontiers of Speech Communication Research. New York: Academic Press.

Klatt, D. H. 1979. Speech perception: a model of acoustic-phonetic analysis and lexical access. Journal of Phonetics. 7:279-312.

Liberman, A. M.; Cooper, F. S.; Shankweiler, D. P. and Studdert-Kennedy, M. 1967. Perception of the speech code. Psychological Review. 74:431-461.

Massaro, D. W. 1972. Preperceptual images, processing time and perceptual units in auditory perception. Psychology Review. 79:124-145.

Morton, J. and Broadbent, D. 1967. Passive versus active recognition models, or is your homunculus really necessary? in Wathen-Dunn, W. (ed) Models for the Perception of Speech and Visual Form. Cambridge, Mass.: MIT Press.

Pisoni, D. and Sawusch, J. R. 1975. Some stages of processing in speech perception. in Cohen, A. and Nooteboom, S. G. (eds) Structure and Process in Speech Perception. New York: Springer-Verlag.

Stevens, K. N. and Halle, M. 1967. Remarks on analysis by synthesis and distinctive features. in Wathen-Dunn, W. (ed) Models for the Perception of Speech and Visual Form. Cambridge, Mass.: MIT Press.

Whalen, D. H. 1984. Subcategorical phonetic mismatches slow phonetic judgments. Perception and Psychophysics. 35:49-64.