

Synchrony capture hypothesis fails to account for effects of amplitude on voicing perception

Andrew J. Lotto^{a)}

Department of Psychology, Washington State University, PO Box 644820, Pullman, Washington 99164

Keith R. Kluender

Department of Psychology, University of Wisconsin—Madison, 1202 West Johnson Street, Madison, Wisconsin 53706

(Received 5 July 2001; revised 8 November 2001; accepted 13 November 2001)

Kluender *et al.* [J. Acoust. Soc. Am. **97**, 2552–2567 (1995)] reported that overall stimulus amplitude affects perception of the voicing contrast in syllable-initial stops as a function of frequency separation between the first formant (F1) and higher formants (F2, F3). These results were offered as support for a hypothesis that [–voice] could be signaled by a shift in the temporal pattern of neural firing from synchronizing to energy at frequencies of F2 and F3 to synchronizing to energy near F1. Several predictions from this “synchrony capture hypothesis” were tested in the current study. In all cases the hypothesis was not supported. Effect of stimulus amplitude (increased voiceless responses with higher amplitude) was maintained when there was no cutback in F1 or when F2 and F1 energy bands were presented dichotically. In further tests of the hypothesis, voice–voiceless series were created that maintained periodic energy throughout the syllable (with F1 cutback signaling voicing). Energy just below the frequency of F2 and energy above F1 were presented dichotically. Thus, at the periphery there was no competition between frequencies near F2 and lower frequencies. In contrast to predictions of the “synchrony capture hypothesis,” overall amplitude still had an effect on voice–voiceless identifications. © 2002 Acoustical Society of America. [DOI: 10.1121/1.1433809]

PACS numbers: 43.71.Es [CWT]

I. INTRODUCTION

The voicing contrast in syllable-initial stop consonants has been a major interest of speech perception researchers for many years. This is due in part to the ubiquity of the voicing distinction in languages (Maddieson, 1984) and in part to the apparent complexity of the mapping from stimulus attributes to perception of the contrast. There are a number of spectral and temporal acoustic attributes that have been shown to affect the identification of a stop as voiced or voiceless (Stevens and Klatt, 1974; Lisker, 1975; Summerfield and Haggard, 1977; Soli, 1983; Kluender and Lotto, 1994).

One attempt to account for the joint effects of some of these spectral and temporal cues is the synchrony capture hypothesis (SCH) offered by Kluender *et al.* (1995). This hypothesis was developed from consideration of the neural response to voiced and voiceless stops and was inspired, in particular, by the work of Sinex and his colleagues on auditory nerve (AN) responses to consonant-vowel (CV) stimuli (Sinex and Geisler, 1983; Sinex and McDonald, 1988, 1989; Sinex *et al.*, 1991). In investigations of synchronization patterns in chinchilla AN fibers, Sinex and MacDonald (1989) demonstrated that mid- and high-CF fibers often responded in synchrony to frequencies near the second formant (F2) of the CV prior to the onset of voicing.¹ These same fibers changed their dominant synchronization to components near the frequency of the first formant (F1) following the onset of

voicing. Kluender *et al.* (1995) proposed that this change in dominant synchronization for a population of fibers at the onset of periodicity was a potential signal for voicelessness. Prior to voicing in *voiceless* stops, there is little energy near F1 and the vocal tract is excited by an aperiodic aspiration source. As a result, mid-CF fibers have an opportunity to synchronize to frequencies near F2 and F3 prior to being “captured” by F1 energy. In contrast, the release of occlusion and onset of voicing in *voiced* stops (in English) tend to be near simultaneous. Thus, there is energy near F1 from the onset of the consonant and no time for mid-CF AN fibers to synchronize to F2 prior to a change in dominant synchronization. Therefore, according to the SCH, *change* in dominant synchronization of a large population of mid- and high-CF fibers is potential information to the voice contrast at the neural level.

Several predictions were derived from this hypothesis based on previous empirical work on neural synchrony (Sachs and Abbas, 1976; Javel *et al.*, 1983). These predictions were tested by presenting listeners series of synthesized CV syllables that varied from voiced to voiceless. The first prediction was that increases in overall amplitude of CVs should result in more stimuli being identified as voiceless. This prediction arises because synchrony capture, the relative dominance of the synchronization by one frequency component over others, increases with increased amplitude (Javel *et al.*, 1983). In addition, increases in intensity lead to an extension of the low-frequency tails of the response area for a neuron. That is, as amplitude increases, the likelihood of a mid-CF fiber being sensitive to (and becoming synchronized to) com-

^{a)}Electronic mail: alotto@wsu.edu

ponents near F1 increases. If this change signals voicelessness, as proposed by the SCH, then higher amplitudes should result in more voiceless identifications. In agreement with this prediction, Kluender *et al.* (1995) reported that voiceless identifications increased monotonically as CV amplitude increased from 40 to 60 to 80 dB.²

The second prediction was that a decrease in the frequency separation of F1 and F2 would enhance the effect of amplitude on voicing judgments. The rationale for this prediction was that if F1 and F2 were closer in frequency, then energy from F1 would be more likely to fall within the low-frequency tails of a larger population of the fibers that originally synchronized to F2 at stimulus onset. At the onset of voicing, this large group of fibers would more likely switch dominant synchronization to F1, resulting in a more robust signal for voicelessness. Kluender *et al.* reported the results of several experiments that upheld this prediction. When F1 onset frequency was increased or F2 frequency decreased (by changing place of articulation of consonant or identity of following vowel) the effect size of amplitude on identifications was increased.

The SCH is appealing because it relates neural representations to behavior and it appears to account for some of the interaction of spectral and temporal attributes in determining the voice contrast judgment. The results of Kluender *et al.* (1995) related above matched novel predictions made by the hypothesis. In addition, Kluender *et al.* ruled out differential masking of aspiration energy as a cause for the effect of amplitude on identifications because the effect of aspiration level was shown not to interact with overall stimulus level. The SCH seems to be well supported.

The purpose of the experiments reported in this article is to further test the hypothesis by examining some additional predictions of the SCH. The studies reviewed above tested the viability of the SCH by manipulating stimulus parameters thought to enhance the likelihood of synchrony capture. The predictions of the hypothesis would be that these manipulations should lead to more voiceless judgments and that these effects would be a function of overall stimulus level. The present experiments attempt to eliminate the possibility of mid-CF fibers changing their dominant synchronization in CV stimuli. This should eliminate synchrony capture as a cue for voicelessness. The prediction of the SCH is *not* that this will lead to the absence of voiceless responses. There are certainly other attributes that can signal voicelessness, and synchrony capture was never meant to be the sole determiner of the perception of the contrast. Instead, the SCH would predict that without major shifts in temporal encoding from F2 to F1, overall stimulus amplitude should have a reduced or nonexistent effect on voice judgments. That is, if the explanation for the amplitude effect is due to synchrony capture, then one should see a substantially reduced effect when changes in dominant synchronization are limited.

II. EXPERIMENT 1: LACK OF F1 CUTBACK

The SCH of Kluender *et al.* (1995) suggests that the *change* in dominant synchronization for mid-CF fibers from components near F2 to components near F1 during a CV is a cue to voicelessness. The reason this change occurs in voice-

less stimuli is because there is little or no energy in the region of F1 prior to voicing (referred to as F1-cutback). One way to eliminate or diminish this potential cue is to present CVs with energy present in the region of F1 throughout the stimulus. With the elimination of F1-cutback most mid-CF fibers should synchronize to F1 from stimulus onset. Thus, there will be a greatly diminished *change* in synchronization to act as a cue. If modulation of the strength of the synchronization cue is the explanation for amplitude effects on voicing judgments, then CV series without F1-cutback should show a smaller identification shift with change in amplitude.

A. Methods

1. Stimuli

Two series of three-formant syllables were created based on the series used in experiment 2 of Kluender *et al.* (1995). One series varied perceptually from /ba/ to /pa/ (labial). The other series varied perceptually from /ga/ to /ka/ (velar). Stimuli were generated using the parallel branch of the Klatt (1980) software synthesizer. The only difference between stimuli in the two series was the frequency of F2 onset. For the labial series, F2 began at 800 Hz, increasing linearly to a steady state of 1220 Hz after 45 ms. For the velar series, F2-onset frequency was 1950 Hz, decreasing linearly to a steady state of 1220 Hz after 45 ms. F1 and F3 parameters were the same for both series. The frequency of F1 at stimulus onset was 300 Hz, rising to 750 Hz over the 45-ms transition duration. F3 began at 1200 Hz, increasing to a steady state of 2600 Hz over the same period. Fundamental frequency (f_0) was 128 Hz at the beginning of the syllable and decreased to 100 Hz over the last 50 ms. Overall stimulus duration was 300 ms.

For each series, there were 13 stimuli differing in the duration between stimulus onset and voicing onset (5–65 ms in 5-ms steps). Prior to the onset of voicing, all formants were excited with a noise source (amplitude of aspiration synthesizer control parameter AH=65). Note that, while Kluender *et al.* (1995) covaried F1-cutback with the aspiration energy, onset of F1 energy and the onset of energy near F2 and F3 are contemporaneous for these stimuli. All stimuli were matched in rms amplitude of the steady-state portion of the vowel.

Stimuli were synthesized with 12-bit resolution at a 10-kHz sampling rate and stored on computer disk. Stimulus presentation and data collection were under the control of a microcomputer. Following D/A conversion (Ariel DSP-16), stimuli were low-pass filtered (Frequency Devices 677, cut-off frequency 4.8 kHz) prior to being attenuated (Analog Devices AD7111 digital attenuator), amplified (Stewart HDA4), and played over headphones (Beyer DT-100). For each stimulus, the rms energy of the steady-state vowel was matched to a 1000-Hz sine wave of 40, 60, or 80 dB SPL (A) intensity. Calibration of the audio presentation was accomplished by using a Brüel & Kjaer system consisting of a flat-plate adapter (type DB0843) on an artificial ear (type 4153) with $\frac{1}{2}$ -in. microphone (type 4134) and sound level meter (type 2203).

2. Subjects

Fourteen college-aged adults participated. All subjects learned English as their first language and reported normal hearing. Each participant received class credit in an Introductory Psychology course.

3. Procedure

One to three subjects were run at one time in three single-subject soundproof chambers (Suttle Equipment Corp.) during a single experimental session consisting of 12-min test periods separated by brief breaks. During each of the three test periods, the two series (labial and velar) of 13 stimuli (5- to 65-ms aspiration noise duration in 5-ms steps) were presented three times in random order at each of the three intensity levels (40, 60, and 80 dB). This yielded a total of nine presentations of each stimulus at each intensity level. Participants were instructed to press a button labeled “BDG” if they heard the syllable as beginning with one of these consonants or to press a button labeled “PTK” to indicate that they perceived one of these consonants. The alveolar responses (D,T) were included in the unlikely case that the participants happened to hear some of the stimulus exemplars as /da/ or /ta/. Thus, responses were generally in terms of voiced versus voiceless.

B. Results and discussion

Kluender *et al.* (1995) did not include in analyses data from any subject who was not able to correctly label the endpoint stimuli (i.e., 5 ms as voiced and 65 ms as voiceless) at least 90% of the time. This was done to exclude participants who may have had difficulty with presentations of syllables at 40 dB. The same criterion was used for inclusion of data in the present experiments. In experiment 1, 6 of 14 subjects failed to respond correctly for at least 90% of the presentations across the 12 end-point stimuli. This ratio of subject loss is slightly higher than that of Kluender *et al.* This may be due in part to the unnaturalness of the stop consonants with no F1-cutback.³

Boundary values for each of the six series (labial/velar × presentation level) were calculated for each subject using probit analysis. The SCH would lead to the prediction that for each series (labial and velar) boundary values should remain fairly constant across presentation level. This is because the cue that is proposed to vary with amplitude, synchrony change, has been diminished or eliminated. The mean boundary for each condition is displayed in Fig. 1. As can be seen, the data do not support the prediction of the SCH. For each series, boundaries shift as a function of presentation level. This shift is similar to that reported by Kluender *et al.*, i.e., more voiceless responses at higher overall amplitudes.

Separate one-way within-subjects ANOVA were run on data for each series. These tests revealed that amplitude-based boundary shifts were evident for both the labial [$F(2,14) = 49.35, p < 0.0001$] and the velar [$F(2,14) = 20.21, p < 0.0005$] series. Calculation of protected least significant difference (Keppel, 1982) revealed that identification boundaries differed for each amplitude level in each series ($\alpha = 0.05$). An analysis of the mean percentage of [+voice] (“BDG”) responses in each condition (as opposed

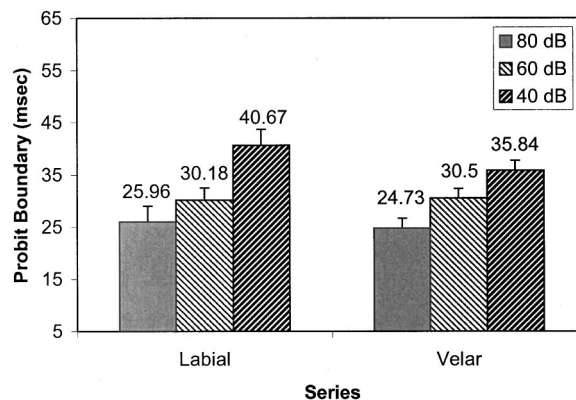


FIG. 1. Mean boundary values (and attendant standard errors) for each series (labial and velar) from experiment 1. The parameter is the overall presentation level (80, 60, or 40 dB).

to the boundaries) provided an equivalent pattern of results. Greater stimulus amplitude led to significantly more “voiceless” responses [$F(2,14) = 52.11, p < 0.0001$]. In agreement with the results of Kluender *et al.*, the effect of amplitude was slightly greater for the labial series [$F(2,14) = 9.01, p < 0.005$ for the interaction between place of articulation and stimulus amplitude]. The data clearly demonstrate that the effect of amplitude on voicing judgments is present even for CV stimuli that lack F1-cutback, in opposition to predictions of the SCH.

A more lenient test of the SCH would be to demonstrate that the effect of amplitude decreases substantially with no F1-cutback even if it does not disappear outright. However, the boundary shifts for these series were comparable to the ones obtained by Kluender *et al.* In the present experiment the mean boundary shift for the labial series was 14.71 ms compared to 12.24 ms for Kluender *et al.* (1995, experiment 2). The respective values for the velar series are 11.11 and 3.77 ms. Thus, there is no support here for the predictions of the SCH.

One way to salvage the SCH given these results is to suggest that despite the presence of energy near F1 during the aspiration portion of the CV, this energy was not sufficiently intense to capture the synchronization of mid-CF fibers. The spectrum for the aspiration source in the Klatt (1980) synthesizer is relatively flat. As a result, the relative amplitude of the F1 peak will not differ as much from the amplitude of the F2 and F3 peaks as when the voicing source is used. Perhaps, this reduced difference in amplitude minimized synchrony to F1 by mid-CF fibers. If this were the case, then temporal responses of these fibers would still be dominated by energy near F2 and F3. At onset of voicing, amplitude of energy near F1 increases and could capture the dominant synchronization of these fibers. That is, despite the lack of F1-cutback, change in synchronization may remain a viable cue to voicing.

III. EXPERIMENT 2: DICHOTIC PRESENTATION

Experiment 1 attempted to eliminate the synchrony change cue to voicing by encouraging neurons to encode frequencies near F1 for the duration of the stimulus. Experi-

ment 2 takes a complementary tack. The stimuli are manipulated to eliminate the possibility of mid-CF fibers synchronizing to F1 frequencies at any time during the duration of the CV syllable. This was accomplished by filtering the stimulus and presenting low-frequency energy around F1 to one ear and energy around F2 and F3 to the opposite ear. Because there is no possible opportunity for AN fibers to shift synchronization from F2 to F1, changing stimulus amplitude cannot make the shift more or less likely. According to the SCH, this lack of amplitude affect on synchronization change should eliminate or greatly reduce the effect of amplitude on voicing judgments.

A. Methods

1. Stimuli

Again, velar and labial series of 13 synthesized CVs were created. These stimuli were identical to the original series used by Kluender *et al.* (1995, experiment 2). Unlike experiment 1, energy in the region of F1 was attenuated during the aspirated portion of the signal as occurs in natural productions. This was accomplished by manipulating the amplitude of the filter associated with F1 (parameter A1 set to 0 at syllable onset and to 55 at voicing onset). Each CV was filtered twice using a sharp (eight-pole) elliptical digital filter. For both filters, the cutoff frequency was 1000 Hz. One filter was high pass, resulting in a filtered CV that contained only the second and higher formants. The other filter was low pass, resulting in a filtered CV containing predominantly F1 information. For velar stimuli, this low-pass filtered stimulus contained no appreciable information about F2 as this formant started at 1950 Hz and decreased to 1200 Hz. For the labial stimuli, there was some F2 energy present in this low-pass filtered CV due to the low F2-onset frequency of these syllables (800 Hz). Given the formant structure of the labial syllables, this residue F2 energy was unavoidable. Other than this case, discrete FFTs demonstrated that the filtering resulted in the desired segregation of spectral energy. Filtering was performed off-line and stimuli were saved to disk. The two filtered portions of the CVs were presented time-aligned to opposite ears at 80, 60, and 40 dB (amplitude of entire stimulus prior to filtering).

2. Subjects

Twenty-two college-age adults participated for Psychology course credit. All reported normal hearing and learned English as a first language.

3. Procedure

The procedure was identical to that used in experiment 1. On each trial, the participants were presented both portions (high and low pass) of the filtered CVs at one of three amplitudes (80, 60, or 40 dB). The low-pass portion was always presented to the left ear. Participants identified the syllable-initial consonant as “BDG” or “PTK.”

B. Results and discussion

Two subjects failed to meet the criterion listed for experiment 1 of 90% correct identification across endpoints.

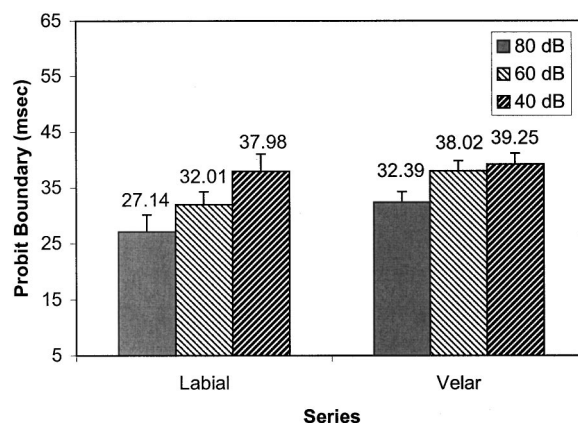


FIG. 2. Mean boundary values (and attendant standard errors) for each series (labial and velar) from experiment 2. The parameter is the overall presentation level (80, 60, or 40 dB).

Their data were not included in the analyses. Probit boundaries were computed from the identification curves of the remaining 20 subjects. Mean boundaries for each series (labial and velar) at each amplitude (80, 60, and 40 dB) are presented in Fig. 2.

A one-way ANOVA was computed on boundary values for each series separately. Boundary values decreased (more voiceless responses) with increases in overall amplitude for both the labial series [$F(2,38) = 111.68, p < 0.0001$] and the velar series [$F(2,38) = 51.12, p < 0.0001$]. *Post hoc* tests revealed that the boundaries differed significantly at all three presentation levels for labial stimuli. For velar stimuli, the identification boundary for 40 dB presentation level differed from both the 80 and 60 dB levels, but the two higher-level conditions did not reliably differ from each other. An equivalent pattern of results is present in the mean percentage of [+voice] (“BDG”) responses. Across place, there is a significant effect of stimulus amplitude [$F(2,38) = 87.71, p < 0.0001$].

These data do not support the predictions of the SCH. Mid-CF fibers in the right ear should have synchronized to the components of F2 and F3 at stimulus onset and there would be no opportunity to change synchronization to F1 because there was no energy near F1 presented to that ear. In the left ear, for velar stimuli, there was little or no energy at stimulus onset followed by energy near F1 at the onset of voicing. Thus, mid-CF fibers would again have no opportunity to shift dominant synchronization. For velar stimuli, there should be no change-in-synchrony cue in the firing patterns of the VIIIth nerve for either ear. The SCH would, therefore, predict no effect of amplitude. However, there was an average boundary shift of about 7 ms for the velar series, which is even larger than that obtained in the original experiments of Kluender *et al.*

The results of experiments 1 and 2 provide little evidence supporting the notion that modulation of the shift in synchronization of AN fibers from F2/F3 to F1 is the basis for the effect of amplitude on voicing judgments. However, an alternative synchronization shift cue may account for these data. At the onset of voicing in a CV syllable, there is not only onset of F1 energy but also onset of energy at the

fundamental frequency. Mid-CF fibers could change their dominant synchronization from F2/F3 to f_0 at voicing onset. This shift could potentially serve as a cue to voicelessness. This slightly modified version of the synchrony capture hypothesis (mod-SCH) would be compatible with the results of experiment 1. The lack of F1 cutback would not affect this shift. What is important for mod-SCH is “ f_0 cutback,” so to speak. The aspirated portion of the syllable would allow AN fibers to fire in synchrony to components near F2/F3. Then, the fibers’ temporal patterns may encode f_0 at voicing onset.

The mod-SCH may also be compatible with the results of experiment 2. Despite the filtering of lower frequencies, f_0 remains present in the high-pass CV stimuli in the harmonics. Theoretically, mid-CF AN fibers in the right ear could first synchronize to F2 and F3 and then switch synchronization to f_0 at onset of voicing. The likelihood of this shift would be dependent on amplitude, with more fibers shifting synchronization at louder presentation levels. Thus, if the change in temporal synchronization were a cue to voicelessness, one would predict more voiceless judgments at higher intensities. This is exactly the pattern of results in experiment 2.⁴ In order to test the SCH-mod, experiment 3 was designed as a variation of experiment 2.

IV. EXPERIMENT 3: DICHOTIC PRESENTATION, NO f_0 CUTBACK

In order to eliminate the possibility of mid-CF fibers *changing* dominant synchronization to f_0 , stimuli were created that signaled voicelessness only through cutback in F1 energy. That is, periodic energy was present throughout the CVs. Absence or presence of energy in the region of F1 remained as a cue to the voicing contrast. These stimuli were then filtered and presented dichotically as in experiment 2. These manipulations should, in tandem, significantly reduce the possibility of sharp changes in synchronization during the CVs. It is clear in this case that the SCH (and the mod-SCH) would predict that the effect of amplitude on voicing judgments should be reduced significantly or should disappear completely.

A. Methods

1. Stimuli

Two series of CVs were synthesized that were identical to those employed in experiment 2 except that periodic energy excited the formants throughout the syllable. That is, there was no aspirated portion in the initial stops (AH=0 throughout). Each step of the series differed in the timing between syllable onset and the onset of significant energy in the region of F1 (controlled by the amplitude of the filter for the first formant; A1 parameter). This duration varied from 5 to 65 ms in 5-ms steps. Each syllable was filtered using the same filters as described in experiment 2. Because high-pass-filtered syllables of each series did not change with varying F1 cutback, the same base labial and velar high-pass-filtered stimuli served as one of the dichotic inputs for each member of the series (the 5-ms F1-cutback stimulus in each case). Thus, the input to this ear provided no information about the voice contrast.

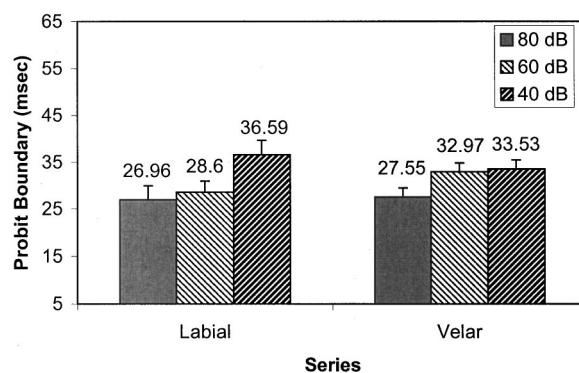


FIG. 3. Mean boundary values (and attendant standard errors) for each series (labial and velar) from experiment 3. The parameter is the overall presentation level (80, 60, or 40 dB).

2. Subjects

Twenty-eight college-aged adults participated in the experiment for class credit. All reported normal hearing and English as their native language. None of the subjects had participated in experiments 1 or 2.

3. Procedure

The procedure was identical to experiment 2. Only the stimuli were changed.

B. Results and discussion

The stimuli used in experiment 3 were very unnatural. The lack of delay in the onset of periodicity eliminates one of the most salient attributes of voiceless stimuli. In addition, the syllables were filtered and presented dichotically; a decidedly unlikely acoustic event outside the laboratory. Not surprisingly, many participants had difficulties consistently labeling these stimuli. What may be surprising is that 25% of the subjects (7 out of 28) were able to label the syllable-initial consonants well enough to pass the rather strict criterion of 90% correct across endpoints. Identification functions from these subjects allowed for the computation of probit boundaries, which can be compared to the results from experiments 1 and 2. Given the large number of subjects failing to reach criterion, both sets of data will be examined separately below.

Mean identification boundaries for each condition were computed from the data of the subjects who passed criterion. Means and standard errors of these boundaries are displayed in Fig. 3. These boundaries were entered into separate one-way ANOVA for each series. The analyses reveal that the effect of amplitude is statistically significant for both the labial [$F(2,18)=5.64, p<0.05$] and velar series [$F(2,18)=6.03, p<0.01$]. According to *post hoc* comparisons, the boundary for the labial series presented at 40 dB differed significantly from both the 80- and 60-dB boundaries, which did not differ from each other. For the velar series, the 80-dB boundary differed from both the 60- and 40-dB boundaries, which did not differ from each other. Again, the effect of amplitude can also be seen in analyses of the mean percentage of [+voice] responses [$F(2,12)=12.15, p<0.005$ across place of articulation].

TABLE I. Mean percentages of [+voice] responses for subjects failing to meet criterion in experiment 3. The atypical data from one subject are presented. Results of one-way ANOVA are included for all subjects failing criterion ($df=2,57$) and for the same group without the data of the atypical subject ($df=2,54$).

	80 dB mean	60 dB mean	40 dB mean	F-statistic	p value
Labial series (all subjects)	56.07	65.22	68.98	3.16	0.097
Velar series (all subjects)	55.91	62.40	66.84	1.31	0.28
Labial series (atypical subject)	94.88	54.71	13.66		
Velar series (atypical subject)	88.05	55.56	10.25		
Labial series (without atypical)	54.03	65.77	71.89	5.37	0.0074
Velar series (without atypical)	54.22	62.76	69.82	2.87	0.065

Because the number of subjects who did not reach criterion was so large, their data was analyzed separately to see if they demonstrated similar trends to those described above. Identification data from these subjects did not provide functions that were amenable to probit analysis. As a result, mean percent [+voice] responses (“BDG” label) were calculated for each subject in each condition. These means were entered into separate one-way ANOVA for each series. The mean [+voice] judgments *increased* as overall amplitude *decreased* for both series. However, these differences did not reach statistical significance (see Table I). In examining the individual data, one subject’s responses were clearly aberrant. This single subject showed a very large effect of amplitude in the opposite direction from the pattern present in the other 20 subjects’ data. When this subject’s data were eliminated from the analyses, the effects of amplitude on voicing judgments reached statistical significance (see Table I).

The data from both groups of subjects paint a fairly clear picture. The effect of amplitude is still present even for CV stimuli that do not offer an opportunity for mid-CF fibers to change dominant synchronization to components of F1 (experiments 1 and 2) or f_0 (experiment 3). There isn’t any evidence that these effects have been weakened by the stimulus manipulations for experiment 3. The original boundary shifts reported by Kluender *et al.* (1995) were 12.24 ms for the labial series and 3.77 ms for the velar series. For experiment 3, the respective shifts were 9.63 and 5.98 ms. The predictions of mod-SCH were not obtained.

V. GENERAL DISCUSSION

The synchrony capture hypothesis offered by Kluender *et al.* (1995) produced novel predictions about human responses to speech sounds based on data from encoding of CVs in the auditory nerve. The fact that SCH predictions were consistent with multiple empirical tests provided an example of the utility of integrating neurophysiological results into speech perception theory. The experiments presented here were designed to further test the validity of this hypothesis. In each experiment, CV stimuli were manipulated to reduce the probability that mid-CF auditory nerve fibers would change their dominant synchronization during the course of the stimulus. According to the SCH, these ma-

nipulations should reduce the shift in voiced–voiceless identifications that occur when overall stimulus amplitude changes. The results of these three experiments were unequivocal. Whether F1 cutback was eliminated or lower and higher frequencies were presented to opposite ears, the effect of amplitude on voicing judgments remained undiminished. Even the extreme manipulation of deleting the delay between stimulus onset and voicing onset (and presenting the stimuli dichotically) did not alter the effect of amplitude.

Absent an explanation from the SCH,⁵ the robust effect of amplitude on voicing judgments as well as the details of effects related to frequency separation between F1 and higher formants remain unexplained. These identification shifts, first described by Van Tassel and Crump (1981), have obvious implications for speech amplification. The size and resiliency of these effects are remarkable when compared to other speech perception effects. Stimulus amplitude effects remained strong through all of the stimulus manipulations of Kluender *et al.* (1995) and the present set of experiments.

One alternative explanation for these shifts is that lower amplitudes compromise the detection of the onset of aspiration. The duration between aspiration onset and voicing onset is considered an important cue to the voicing distinction (Lisker and Abramson, 1964). Cortical potentials elicited from stop+ vowel stimuli usually show a peak corresponding to aspiration or burst onset and a second peak corresponding to the onset of voicing (Koch *et al.*, 1997; Sharma and Dorman, 1999). It is possible that when the amplitude of the syllable is decreased, the neural representation of the aperiodic energy onset is more affected than the onset of voicing. Periodic energy in natural speech is typically higher in amplitude [it was in all of the stimuli used by Kluender *et al.* (1995)] and the predictable nature of the periodic signal may provide some protection from the effects of lower amplitude. In contrast, the onset of aspiration energy may be easily compromised. From a signal detection viewpoint, more temporal information may be required to detect the onset of the aspiration noise when the amplitude is decreased. This could lead to more [+voice] responses at lower amplitudes, which matches the effect as reported (Van Tassel and Crump, 1981; Kluender *et al.*, 1995). In addition, this explanation would predict less of an effect of amplitude for velar stimuli. Velar

consonants are distinguished by a compact concentration of aspiration energy as the frequencies of F2 and F3 begin close together. This should result in a clearer marking of the onset of the aspiration in those frequency bands. As a result, the deleterious effects of amplitude would be ameliorated. This prediction is supported in the current data and in data from Kluender *et al.*

As plausible as this explanation may seem, it also fails to account for all of the data. In experiment 4 of Kluender *et al.*, aspiration level was manipulated but did not interact with the effect of overall stimulus amplitude. In their experiment 5, natural tokens were used with bursts present or filtered off. Bursts should provide a compelling signal to stimulus onset, but burst presence did not modulate the effects of amplitude. In experiment 3 of this report, the voice contrast was signaled only by F1-cutback with periodic energy present throughout the stimulus. The resulting identification functions still showed boundary shifts that were dependent on amplitude. These three stimulus manipulations should have provided better markers for stimulus onset, yet none appeared to interact with overall amplitude.

The effect of amplitude on voicing judgments remains unexplained. Synchrony capture cannot explain the effect. While the present results most clearly eliminate synchrony capture as an explanation for effects of amplitude, it must be noted that the only perception data supporting the SCH relied critically upon amplitude effects being closely tied to changes in synchrony. As such, the SCH is left to rest solely upon physiological observations from cat and chinchilla. Absent compelling new data from humans, the SCH must be viewed with some skepticism.

ACKNOWLEDGMENTS

Some of the data were presented at the 131st Meeting of the Acoustical Society of America in Indianapolis, IN. The authors thank Quentin Summerfield for comments on the original synchrony capture paper that inspired some of these experiments. Research supported by NSF Grant No. DBS-9258482.

¹We will use the term “voicing” to refer to the presence of periodic energy. This is somewhat imprecise because the term is articulatory and we are only concerned here with attributes of the acoustic signal regardless of the source. In addition, we use the terms “labial” and “velar” to refer to stimuli differing in F2 onset frequency. However, this manner of usage is common in the speech field and we have decided to sacrifice precision for ease of reading. We have avoided the articulatory term “voice-onset time” because its use has traditionally confounded several attributes of the signal (e.g., F1-cutback, duration of aspiration) that are manipulated independently in this study.

²The steady-state vowel portions of these stimuli were matched in rms energy to a 1000-Hz sine wave of 40, 60, or 80 dB intensity.

³The trends described in the results section were also obviously present in the data of the subjects who failed to pass criterion.

⁴Kluender *et al.* (1995) concentrate on synchrony capture by F1 and generally ignore f_0 . Part of the reason for this is that the computational models on which the SCH are based (Jenison *et al.*, 1991) use a synchrony measure that is fairly insensitive to f_0 .

⁵Here we are dismissing the hypothesis that modulation of synchrony capture by F1 (or f_0) underlies the effects of amplitude on voicing judgments. We are *not* claiming that synchrony capture does not exist in the auditory nerve nor that synchrony capture has no effect on perception of auditory signals.

- Javel, E., McGee, J., Walsh, E. J., Farley, G. R., and Gorga, M. P. (1983). “Suppression of auditory nerve responses. II. Suppression threshold and growth, iso-suppression contours,” *J. Acoust. Soc. Am.* **74**, 801–813.
- Jenison, R. L., Greenberg, S., Kluender, K. R., and Rhode, W. S. (1991). “A composite model of the auditory periphery for the processing of speech based on the filter response functions of single auditory-nerve fibers,” *J. Acoust. Soc. Am.* **90**, 773–786.
- Keppel, G. (1982). *Design & Analysis* (Prentice-Hall, Englewood Cliffs, NJ).
- Klatt, D. H. (1980). “Software for a cascade/parallel formant synthesizer,” *J. Acoust. Soc. Am.* **67**, 971–995.
- Kluender, K. R., and Lotto, A. J. (1994). “Effects of first formant onset frequency on [-voice] judgments result from auditory processes not specific to humans,” *J. Acoust. Soc. Am.* **95**, 1044–1052.
- Kluender, K. R., Lotto, A. J., and Jenison, R. L. (1995). “Perception of voicing for syllable-initial stops at different intensities: Does synchrony capture signal voiceless stop consonants?,” *J. Acoust. Soc. Am.* **97**, 2552–2567.
- Koch, D., Tremblay, K., Dunn, I., Dinces, E., Carrell, T., and Kraus, N. (1997). “Speech-evoked N1 and mismatch neurophysiologic responses in cochlear implant users and normal listeners,” *Assoc. Res. Otolaryngol. Abstr.* **20**, 80.
- Lisker, L. (1975). “Is it VOT or a first-formant transition detector?,” *J. Acoust. Soc. Am.* **57**, 1547–1551.
- Lisker, L., and Abramson, A. S. (1964). “A cross-language study of voicing in initial stops: Acoustical measurements,” *Word* **20**, 384–422.
- Maddieson, I. (1984). *Patterns of Sounds* (Cambridge U.P., Cambridge).
- Sachs, M. B., and Abbas, P. J. (1976). “Phenomenological model for two-tone suppression,” *J. Acoust. Soc. Am.* **60**, 1157–1163.
- Sharma, A., and Dorman, M. (1999). “Cortical auditory evoked potential correlates of categorical perception of voice-onset-time,” *J. Acoust. Soc. Am.* **106**, 1078–1083.
- Sinex, D. G., and Geisler, C. D. (1983). “Responses of auditory-nerve fibers to consonant-vowel syllables,” *J. Acoust. Soc. Am.* **73**, 602–615.
- Sinex, D. G., and McDonald, L. P. (1988). “Average discharge rate representation of voice-onset time in the chinchilla auditory nerve,” *J. Acoust. Soc. Am.* **83**, 1817–1827.
- Sinex, D. G., and McDonald, L. P. (1989). “Synchronized discharge rate representation of voice-onset time in the chinchilla auditory nerve,” *J. Acoust. Soc. Am.* **85**, 1995–2004.
- Sinex, D. G., McDonald, L. P., and Mott, J. B. (1991). “Neural correlates of nonmonotonic temporal acuity for voice onset time,” *J. Acoust. Soc. Am.* **90**, 2441–2449.
- Soli, S. (1983). “The role of spectral cues in discrimination of voice onset time differences,” *J. Acoust. Soc. Am.* **73**, 2150–2165.
- Stevens, K. N., and Klatt, D. H. (1974). “Role of formant transitions in the voiced-voiceless distinction for stops,” *J. Acoust. Soc. Am.* **55**, 653–659.
- Summerfield, Q., and Haggard, M. (1977). “On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants,” *J. Acoust. Soc. Am.* **62**, 435–448.
- Van Tassel, D. J., and Crump, E. S. A. (1981). “Effects of stimulus level on perception of two acoustic cues in speech,” *J. Acoust. Soc. Am.* **70**, 1527–1529.