In *Proceedings of the 9th International Symposium on Hearing*, 9-14 June 1991, Carcens, France.

Identification of Initial Stop Consonants Processed by the Patterson-Holdsworth ASP Model[1]

Robert Allen Fox and Lawrence L. Feth

Division of Speech and Hearing Science
The Ohio State University
Columbus, OH U.S.A. 43210-1002

Our research group[2] has been developing a model of speech perception which incorporates the Patterson & Holdsworth (1990) Auditory Sensation Processing (ASP) model as a front-end. The present study is aimed at examining the types of dynamic auditory cues that may be important in the perception of syllable-initial stop consonants. This research is a natural extension of the work done in the past decade searching for invariant, possibly time-varying, acoustic characteristics of initial consonants in English (e.g., Blumstein & Stevens, 1979, 1980; Lahiri & Blumstein, 1981; Lahiri, Gewirth, & Blumstein, 1984; Ohde & Stevens, 1983; Kewley-Port, 1983; Kewley-Port & Luce, 1984; Kurowski & Blumstein, 1987). The existence of such cues would seriously compromise theories of speech perception such as the revised Motor Theory (Liberman & Mattingly, 1985) and the Direct Realist Model (Fowler, 1986, 1989; Fowler & Rosenblum, 1991) which assume that unique, possibly innate, linguistic modules are required to identify linguistic segments since segments may not be processed on the basis of contextually varying acoustic signals.

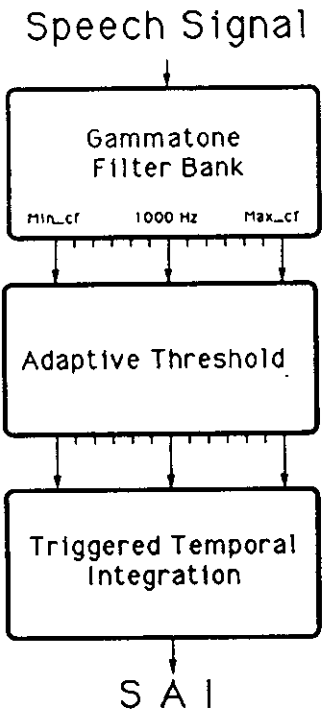
Most recently, several researchers have attempted to incorporate auditory processing algorithms in the analysis of the speech tokens in order to better model human speech perception. It may be the case that the human perceptual system may reduce, in some way, the variability of the acoustic signal. For example, in a study representative of this approach, Kewley-Port & Luce (1984) searched for the time-varying "auditory" features which signaled place of articulation in initial voiced and voiceless stops. To incorporate auditory processing Kewley-Port & Luce required subjects to examine "auditory filtered" running spectral displays which were smoothed spectral sections (frames) calculated and updated every 5 msec. The smoothed LPC spectra were convolved with a bank of 1/6 octave bandwidth filters and the scale of the resulting spectra were changed to a technical mel scale. Two subjects were then required to locate the burst frame and the onset of voicing from each running spectral display and to judge spectral tilt and the presence of midfrequency peaks and to assign phonetic labels to the onset consonants on the basis of these features. Although their subjects achieved an 86% level of accuracy, one may suggest that their auditory representations were not an adequate or reasonable reflection of actual dynamic auditory processing. The 1/6 octave filter bank and mel-scale frequency axis were used to approximate the peripheral filtering (frequency selectivity) of the normal ear. However, the subjects were asked to make their decisions on the basis of physical features selected from the running spectral displays.

The ASP front-end used in our perceptual model more closely simulates the peripheral filtering of the human auditory system. Further, adaptive thresholding and non-linear feature enhancements, now thought to characterize the signal-driven processing of auditory inputs, is also incorporated in the Patterson-Holdsworth model. A schematic representation of this model is shown in Figure 1. Rather than processed, running spectra, the output of the ASP model is a Stabilised Auditory Image (SAI) which preserves the spectral dynamics of time-varying, complex sounds in a

[1] This research was partially supported by grants from the Cognitive Science Center at The Ohio State University, Air Force grant AFOSR-89-0227 (to L. Feth) and NIA grant #5R01 AG08353-02 (to R. Fox).

[2] The authors together with our colleagues Stan Ahalt, Mary Beckman and Ashok Krisnamurthy at Ohio State and Roy Patterson at MRC-APU, Cambridge University.

Figure 1. Schematic of ASP model.



display that mimics the assumed pattern of excitation within the nervous system.

To evaluate the efficacy of this perceptual model, an identification study based on visual cues (the SAI) and broadly similar to the Kewley-Port and Luce (1984) study was designed. The specific goal of this study was to determine whether a sufficient set of the dynamic cues were available in the SAI and recoverable by viewers which could signal both place-of-articulation and voicing distinctions of initial stop consonants in American English. Our long-term goal is to use such cues in the SAI in the development of a model of speech perception (recognition).

Method

Stimuli. A set of CVC monosyllables were recorded. These monosyllables were composed of all combinations of six initial consonants [b d g p t k], five medial vowels [ɪ ɛ æ ɑ u] and three final consonants [b d g]. These manipulations allowed variation in place of articulation and voicing distinctions in a representative set of vowel contexts. A male talker (a trained phonetician) produced six different version of these 30 tokens in the neutral phonetic context *this is a* _____ for a total of 180 different tokens. These tokens were then low-pass filtered at 4.5 kHz, digitized at a 10k sampling rate with 16-bit quantization, and stored on computer disk. Each stimulus token was created by waveform editing the original production. Each token began 15 ms before stop closure release. All initial voiced stops contained prevoicing. The final stop consonant for each token was eliminated and the token shortened to a duration of 200 ms. Onsets and offsets were linearly ramped (5 ms duration) from/to zero. The stimulus set was thus comprised of 180 CVCs beginning with one of six initial consonants ([b d g p t k]).

Table 1. ASP parameter values used in creating the SAIs.

Parameter	Value	Parameter	Value
window width	300 pixels	compression	on
window height	480 pixels	nonlinear saturation	off
magnification	6	upward time constant	0.5 ms
token duration	150 ms	no. integration stages	1
start	0 ms	display update period	4 ms
quantization level	16 bits	trigger threshold decay	
sample rate	10,000 Hz	time constant	8 ms
audiogram equalization	on	max. sustained firing rate	80 Hz
min. channel freq.	200 Hz	min. sustained firing rate	20 Hz
max. channel freq.	4400 Hz	NAP decay time constant	24 ms
filter density	2	SAI decay time constant	24 ms
min. filter bandwidth	24.7 Hz	display duration	15 ms
filter quality factor	9.265	full wave rectification	off
filter output gain	4		

Stabilized auditory images for each of the 180 stimulus tokens were created using the parameter values shown in Table 1. The window width (300 pixels) of each image allowed two images to be shown on the computer monitor at the same time. Each frame of the SAI "cartoon" displayed 15 ms of the processed signal, and the overlap between adjacent frames was 4 ms. The duration of the complete SAI was limited to 150 ms—50 ms shorter than the actual token played to subjects during the review process. The 50 ms section eliminated for each token corresponded only to the

relatively steady-state vowel and provided no dynamic cues to initial stop identity.

Three frames from the SAI "cartoons" for two different tokens—[bɑ] and [dɑ]—are shown in Figure 2. Each line in the figure corresponds to the activity level of a different auditory channel. The frames shown in Figure 2a represent an auditory representation of the two tokens shortly after release of the stop closure. The auditory "activity" shown is in response to the prevoicing of the stop, the stop release, and the onset of formant transitions. As can be seen, the auditory activity for the [d] is somewhat higher in the frequency domain. The frames shown in Figure 2b correspond to early portions of the formant transitions with little or no burst information and no steady-state vowel information. Figure 2c corresponds to a later portion of the consonant transition. Note the differences between Figure 2b and 2c (resulting from the falling F2 transition of [gɑ] and the rising transition of [dɑ]).

Half of the stimulus tokens (90) were used in training the subjects on the identification task and the preliminary viewing tests. The remaining 90 tokens were used to test subjects' ability to generalize their identification skills to a new set of tokens following training.

Training Procedure. Subjects viewed "cartoons" of the stabilized auditory images (SAI) on a high resolution computer monitor (Sun 4 SPARC-SLC station). These images are dynamic in nature and "move" in time (representing the time-varying changes which the auditory system undergoes during speech perception). During our talk, we will provide a video-taped demonstration of a representative sample of these SAIs. The basic goals in the training procedure included (1) acquainting subjects with the basic nature of the stabilized auditory image for individual consonants in a controlled fashion; (2) allowing subjects to associate particular visual displays with a given auditory output; and (3) requiring subjects to develop their own criteria for making phonetic judgments based on the visual displays. The third goal was particularly important to this study. Unlike the approach of Kewley-Port & Luce (1984), although individual subjects were familiar with the nature of the stabilized auditory image (and had a basic understanding of the ASP model), we did not provide explicit "rules" for segment identification. Rather, each subject was required to develop her own set of rules associating visual SAI cues with initial stop consonant distinctions.

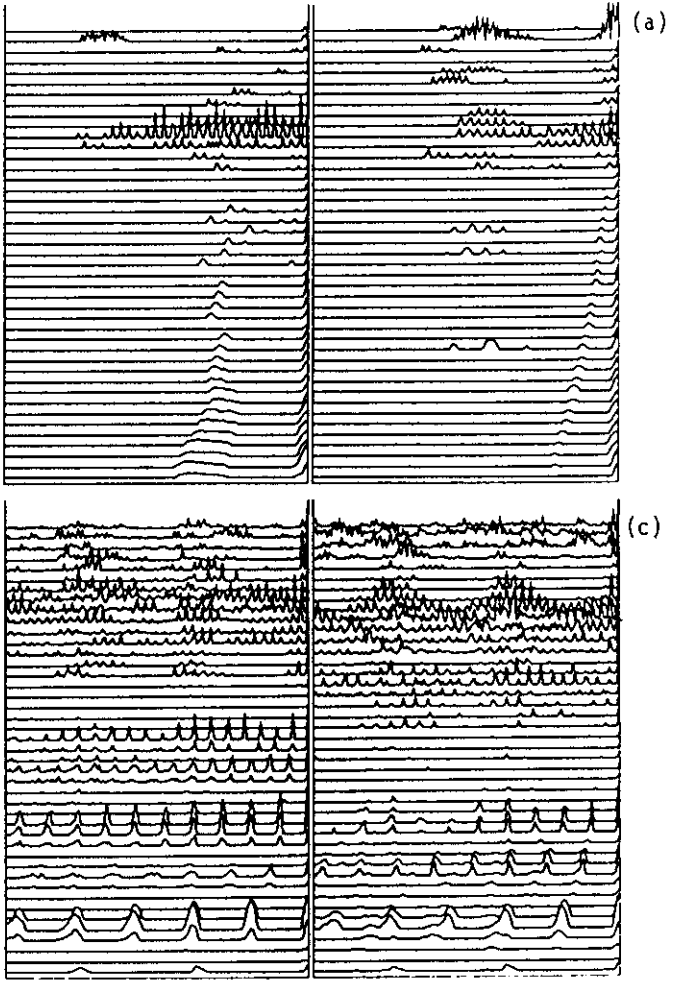
Each subject trained on identifying the stop consonants one vowel quality at a time. That is, at any point in time, a subject viewed the SAIs for a given subset of the tokens and each token in the subset had the same vowel. During the initial viewing of a given set of stops, subjects viewed the stabilized auditory images using a program which both displayed the image and played the acoustic signals through headphones. Later, subjects viewed these images without the accompanying acoustic signal. Subjects were free to examine the SAIs at normal speed (approximately 30 frames/second), at slower speeds (variable and under subjects' control), as well as in a step-wise manner (i.e., one frame of the SAI "cartoon" at a time). Subjects were also asked to compare pairs of tokens during the training session in order to explicitly contrast voicing cues (e.g., [bɑ] vs. [pɑ]) and place cues (e.g., [bɑ] vs. [dɑ], [tɑ] vs. [kɑ]). This comparison was done by presenting two different SAI images side by side on the computer monitor, simultaneously.

In order to discover the specific cues extracted from the SAI by subjects and utilized during the identification tests (see below), subjects were asked to record all impressions, observed visual cues, identification rules, etc., in notebooks which they were allowed to use throughout the training process and during the identification tests. To supplement their observations, subjects were also allowed to make a copy of individual frames of the SAI if the frame illustrated a particularly salient visual cue. This print option was seldom used. Subjects devoted 20 hours/week to the training tasks. These hours were distributed across the entire week—precise distribution of training times was dependent upon the specific schedule of an individual listener. Training on a given set of tokens (grouped by vowel quality) was considered finished only after the subject had individually viewed all 15 tokens, and had made all useful pairwise comparisons. All training and testing (see below) was done on an individual basis with no between-subject discussion in terms of cues utilized. The three subjects trained on the 5 different token subsets in different orders.

Training Pretests. Consonant identification tests were conducted throughout the training period to evaluate the success of the training procedure. Following training on a given subset of tokens (each token having the same vowel quality), a subject was required to identify the initial consonant of a set of 12 randomly selected test tokens taken from that training subset (this constituted the *single-vowel pretest*). These test tokens were identified on the computer disk by random file names, and the subject was required to view each SAI individually and to identify its initial consonant ([b d g p t k]).

Once a subject had trained on tokens from more than a single-vowel set, the subject completed an identification test including CVs from more than a single-vowel group (this constituted the *multiple-vowel pretest*). This pretest forced subjects to rely on vowel-independent cues, when possible, or to identify vowel-specific cue distinctions. No feedback was given during any of the identification tests except for the subject's percentage correct score.

Figure 2. Shown are three frames of the tokens [bɑ] and [dɑ] side-by-side (as subjects see them in the pairwise comparison mode); 2a represents the earliest portion of the stop release; 2b an early part of the consonant transition; 2c a later portion of the consonant transition.



Generalization Testing. The final consonant identification test required subjects to identify the initial consonant of the 90 tokens not used in the training period. Subjects had not viewed these particular tokens before and had not heard them. This final test will reveal the extent to which the auditory cues selected by the subjects are generalizable across other productions of the same CVs by the same speaker. These 90 tokens were provided with random file names and subjects viewed each SAI individually (i.e., no dual displays). Subjects were allowed as much time as desired on each token and were allowed to use both single-step and continuous displays of the SAIs. Subjects were not allowed to use pairwise comparisons.

Results and Discussion

In general, subjects were trainable on the task and were able to identify initial consonants consistently. As expected, voicing distinctions were easier to make than place distinctions. Shown in Figures 3 and 4 are the results of the single-vowel pretests. Subjects correctly identified both place-of-articulation and voicing distinctions 100% of the time in 11 of the 15 single-vowel pretests. Some variation between subjects is evident (e.g., compare subjects 1 and 3) but the worst score for any subject for distinction on any pretest was 83.7% correct. These results show that identification of stop consonants based only on SAI displays is possible, at least when vowel quality is known and unvarying and when subjects are familiar with the tokens.

Figure 3. Percentage correct identification scores of single-vowel pretests for place-of-articulation. Chance level of identification is indicated (here and in all figures to follow) by the horizontal line.

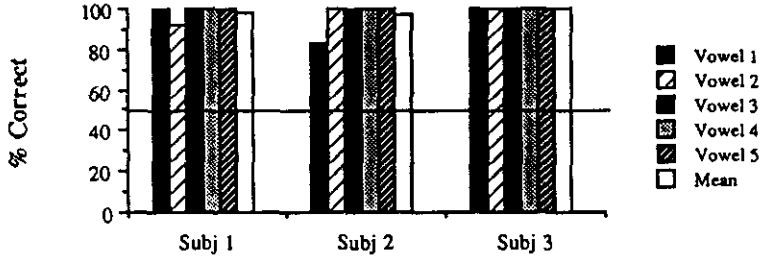
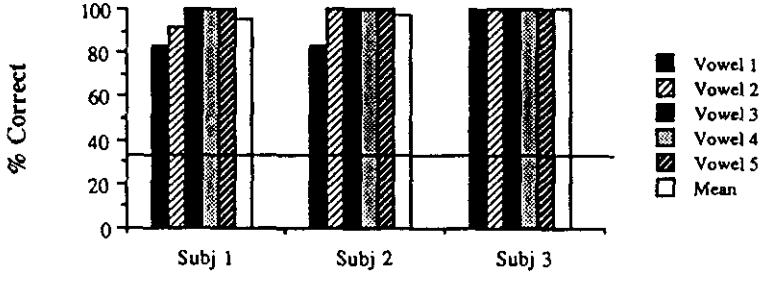


Figure 4. Percentage correct identification scores of single-vowel pretests for voicing.



The multiple-vowel test introduces uncertainty as to the vowel quality of the tokens. If the salient perceptual cues in the SAI are vowel-dependent (especially for place-of-articulation), scores on the multiple-vowel pretests may be systematically lower. Shown in Figures 5 and 6 are the results from the multiple-vowel pretests. It should be noted that the 5-vowel pretests (the rightmost column for each subject) required subjects to identify the initial consonant of all 90 training tokens. Subjects showed perfect identification of voicing in practically all pretests (10 out of 12) with the worst voicing score being 97.8% correct. Identification scores for place-of-articulation were only slightly lower in the multiple-vowel condition for subjects 1 and 3 (97.6 and 95.6%, respectively). Place scores for subject 2 were somewhat lower (83.8%). Subject variations could arise from their use of a significantly different set of cues in the SAI for consonant identification (some sets being more salient than others). However, subject 2 expressed fatigue with the task rel-

atively early and her lowered score could stem from personal sources. In general these pretest scores provide only limited support for the use of vowel-dependent SAI cues. However, other data provide evidence that cues were vowel-dependent. Subject 3 consistently identified (correctly) the entire CV syllable during the pretest. Subject 1 included a description of the SAI cues she used to identify vowel quality in her logbook and her consonantal cues were classified in terms of vowel quality. Verbally, all subjects noted the importance of vowel identity in making their decisions about the initial stop consonant.

Figure 5. Percentage correct identification scores of multiple-vowel pretests for place-of-articulation.

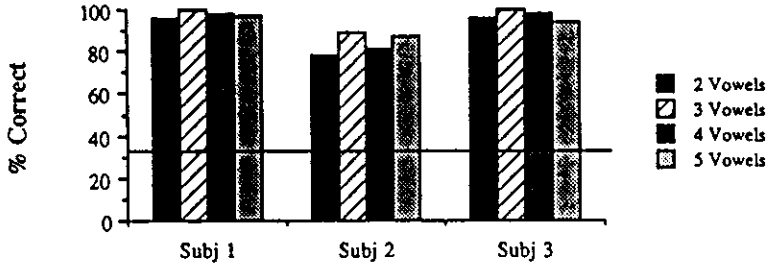
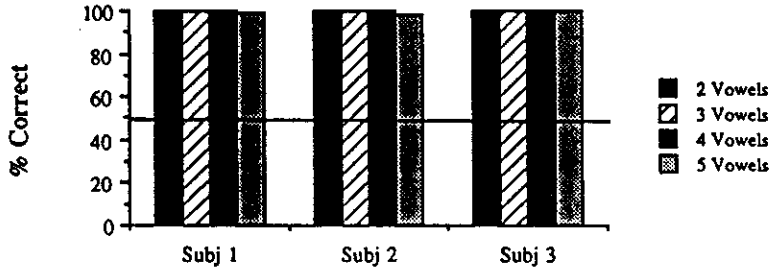
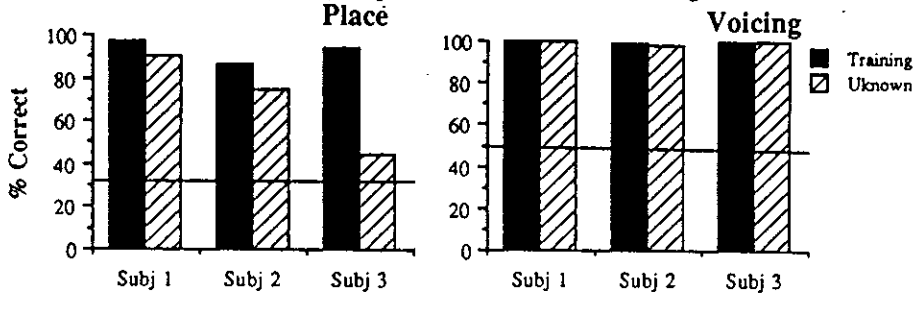


Figure 6. Percentage correct identification scores of multiple-vowel pretests for voicing.



One limitation in evaluating the pretest results is that subjects are identifying only those tokens upon which they trained. It is possible, for example, that during the training sessions subjects became so well acquainted with the training tokens that they were making their identifications based on memory of specific tokens. In the final test, subjects were required to identify a new set of 90 tokens spoken by the same speaker.

Figure 7. Percentage correct identification scores of mean multiple-vowel pretests ("training"—filled columns) compared with scores of final generalization tests ("unknown"—cross-hatched columns) for place-of-articulation and voicing.



Shown in Figure 7 are the mean correct scores on the multiple-vowel tests and the scores for the final test. The difference between the two columns for each subject is an indication of how well the decision criteria developed for the training tokens carried over into the unknown tokens. For all three subjects it is obvious that there is no decrement in ability to discrimination voiced from voiceless tokens—all voicing scores are near 100% correct. A different pattern emerges for the place-of-articulation identifications.

The final place scores for subjects 1 and 2 showed a decrement of about 8% in the final test compared to their multiple-vowel scores. This demonstrated that the cues developed during the training sessions and used in the pretests were broadly applicable to the basic phonetic categories present in the talker's speech and not strictly bound to the first set of 90 tokens. On the other hand, the place scores for subject 3 (whose scores were the best, overall, in the pretests) fell almost to chance level.

In her pretests, subject 3 identified not only the stop consonant as well as the vowel, but name of the specific stimulus token (e.g., ba21 or ga22). An examination of her logbook showed that she extracted very specific and detailed cues for each individual token that sometimes bore little relation to known acoustic dynamics of the sound categories involved. Clearly this subject did not develop broader category cues that could be used across a wider range of the talker's speech. This listener's responses will likely yield little information about the cross-token dynamic auditory feature important to consonant (or vowel) identification but they do point out a danger of the training procedure used here (which did not require listeners to identify unknown tokens until the end of the training) which will be noted in future studies.

Subject 1 and 2 demonstrate that subjects can consistently identify even novel tokens on the basis of the SAI using cues which are not token-dependent but are valid across other productions of the same syllable (these cues could be speaker dependent, but this is a question to be addressed in a later study). At this point we have only done a preliminary analysis of subject 1's logbook to determine the nature of the SAI cues used during the identification tests. In reviewing these cues one should bear in mind that the SAI cartoons did not provide a labeled frequency axis for subjects to use, nor any explicit time scale. Subjects knew, in general, what the SAIs represented, but we provided few measurement "landmarks" for them to utilize. We have provided corresponding acoustic cues commonly noted in the research literature in parentheses whenever possible.

Subject 1's cues for vowel quality included the amount of activity present in the frequency range that would correspond to the second and third formants. Of concern was not only the position of this activity (related to frequency) but the spread of this activity (bandwidth). In the frequency range related to first formant, her cues involve the presence and/or strength of the second and third harmonics (in this range ASP resolves individual harmonics and the frequency range used in creating the SAIs did not allow display of the first harmonic). The vowel [ɪ] was characterized by a strong second harmonic but a "reduced" third harmonic (F0 was approximately 150 Hz, so this SAI pattern would be expected for a first formant around 380 Hz). The vowel [ɑ] was characterized by a second formant in "the middle of the image" (corresponding to about 1200 - 1400 Hz) with strong second and third harmonics (corresponding to a high F1). The vowel [u] was characterized by a 2nd formant in the "middle of the range" with only the second harmonic present (low F1 and mid to low F2). For the vowels [æ ɛ], she only noted the second and third harmonics were "equally strong".

Cues for consonant identification were separated in terms of whether the consonant was voiced or voiceless. Voiced consonants followed by [ɪ ɛ æ] were identified as [g] if "2nd and 3rd formant [were] blended" (i.e., presence of a "velar pinch") and as [b] if there was activity in the lower frequency area at about the 5th frame (onset of the transition—this corresponded to a rising F1 or F2 transition). The consonant was identified a [d] otherwise. Consonants preceding the vowel [ɑ] were identified as [b] if the "2nd formant arise in place and stay there" (rising F2 transition); as [d] if "2nd formant moves down to final spot" (falling F2) and as [g] if there are "blended formants" which can be seen with a running SAI cartoon (velar pinch). Consonants preceding the vowel [u] were identified with respect to SAI activity corresponding to movement of the F2 compared to F3.

Cues for the voiceless consonants included burst information. For example, for syllables with [ɪ ɛ æ] a [p] was characterized by a "small burst—maybe just a few peaks" in lines 18, 19 and 20 (corresponding to mid-frequency burst), the burst for [k] was near the "blended formants ... in the upper 12-13 lines"; [t] often lacked a burst. For the vowel [ɑ], a [k] was characterized by "blended formant"; [t] has a lot of "highest" burst activity (high frequency release spike) and second formant moves down; [p] has a rising 2nd formant.

In summary, these data have demonstrated that subjects can accurately identify stop consonants (and vowels) based on dynamic auditory cues as represented by the stabilized auditory image "cartoons". Subjects can develop (and accurately describe) the cues available in these SAIs which can be related to known acoustic characteristics of the CV syllable used. If these cues could be accu-

rately described, it is possible that they may be used in a model of speech perception or to improve the performance of automatic speech recognition systems. Clearly, it will be necessary to determine the extent to which the dynamic images used here might provide more relevant information for consonant and vowel identification than do more traditional "acoustic" representations (e.g., spectrograms or even the "stabilized auditory spectrograms" derived from SAIs—see Patterson, et al., 1991). However, at this point we are encouraged that these dynamic auditory cues may be successfully utilized in such applications.

References

Blumstein, S. & Stevens, K. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66,* 1001-1017.

Blumstein, S. & Stevens, K. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America, 67,* 648-662.

Fowler, C. (1986). An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics, 14,* 3-28.

Fowler, C. (1989). Real objects of perception. *Ecological Psychology, 1,* 145-160.

Fowler, C. & Rosenblum, D. (1989). In I.G. Mattingly and M. Studdert-Kennedy (eds), *Modularity and the Motor Theory of Speech Perception,* Hillsdale, NJ: Lawrence Erlbaum.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73,* 322-335.

Kewley-Port, D. & Luce, P. (1984). Time-varying features as correlates of place of articulation in stop consonants. *Perception & Psychophysics, 35,* 353-360.

Kurowski, K. & Blumstein, S. (1987). Acoustic properties for place of articulation in nasal consonants. *Journal of the Acoustical Society of America, 81,* 1917-1927.

Lahiri, A. & Blumstein. S. (1981). A reconsideration of acoustic invariance in stop consonants: Evidence from cross-language studies. *Journal of the Acoustical Society of America, Supplement 1, 70,* S39.

Lahiri, A., Gewirth, L. & Blumstein, S. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of the Acoustical Society of America, 76,* 391-404.

Liberman, A. & Mattingly, I. (1985). Motor theory revised. *Cognition, 21,* 1-36.

Patterson, R.D. & Holdsworth, J. (1990). *Technical Report, MRC-Applied Psychology Laboratory,* Cambridge University.

Patterson, R.D., Holdsworth, J., Thurston, P. and Robinson, T. (1991). Auditory images as input for speech recognition systems. To be presented to the *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 19, New Platz, NY.