Perception & Psychophysics
1985, 38, 81-90

Some perceptual properties of consonants in multitalker babble

SANDRA GORDON-SALANT
University of Maryland, College Park, Maryland

This study investigated whether consonant phonetic features or consonant acoustic properties more appropriately describe perceptual confusions among speech stimuli in multitalker babble backgrounds. Ten normal-hearing subjects identified 19 consonants, each paired with /a/, /i/, and /u/ in a CV format. The stimuli were presented in quiet and in three levels of babble. Multidimensional scaling analyses of the confusion data retrieved stimulus dimensions corresponding to consonant acoustic parameters. The acoustic dimensions identified were: periodicity/burst onset, friction duration, consonant-vowel ratio, second formant transition slope, and first formant transition onset. These findings are comparable to previous reports of acoustic effects observed in white-noise conditions, and support the theory that acoustic characteristics are the relevant perceptual properties of speech in noise conditions. Perceptual effects of vowel context and level of the babble also were observed. These condition effects contrast with those previously reported for white-noise interference, and are attributed to direct masking of the low-frequency acoustic cues in the nonsense syllables by the low-frequency spectrum of the babble.

There is a lack of consensus on the relevant perceptual properties of speech that can account for consonant recognition performance. One common approach has been to assess recognition of consonants presented in white-noise backgrounds, to determine the important phonetic features that account for perceptual relations among confused stimuli. However, a single feature-based system has not emerged to account for observed findings across studies. Alternatively, at least one study (Soli & Arabie, 1979) has attempted to explain confusions among stimuli in white noise according to stimulus acoustic properties. However, these limited data preclude any generalizations because findings may be specific to the stimulus set and form of degradation employed. To demonstrate which representation of speech sounds, phonetic or acoustic, best accounts for observed confusions among consonants, the present study examined perceptual effects in a set of listening conditions different from any previously examined.

A variety of experimental paradigms have been used to document the interference effects of white noise. Several important methodological constraints have varied across these studies, which have led to slight differences in the reported results. For example, the speech stimuli utilized have included consonants paired with /a/ in a consonant-vowel (CV) sequence (Miller & Nicely, 1955), consonants paired with /a/, /i/, and /u/ in a CV and vowel-consonant (VC) format (Wang & Bilger, 1973), initial consonants embedded in a monosyllabic word stem (Mitchell & Singh, 1974), and the Modified Rhyme Test (MRT) (Horii, House, & Hughes, 1970). The listening judgment and response tasks have also varied, and include consonant identification (Miller & Nicely, 1955; Wang & Bilger, 1973), word identification (Horii et al., 1970), and similarity judgments via triadic comparisons (Mitchell & Singh, 1974; Singh, 1971). Finally, different methods of analyzing listeners' responses have been applied to these data. One approach has been to analyze performance according to a set of previously defined distinctive phonetic features (Horii et al., 1970; Miller & Nicely, 1955; Wang & Bilger, 1973). An alternative has been to use multidimensional scaling algorithms to extract natural dimensions, and then to identify these dimensions according to the most obvious distinctive phonetic features to emerge (Mitchell & Singh, 1974; Singh, 1971). The latter method has also been applied with an acoustic, rather than a phonetic, feature interpretation of the results (Soli & Arabie, 1979).

Despite the diversity of data collection and analysis methods, certain consistent trends have been observed when the effects of white noise on consonant perception have been studied. Distinctive phonetic feature analyses (both a priori and a posteriori) have generally shown that the relative importance of voicing and nasality increases in white noise, whereas the importance of place and frication decreases (Miller & Nicely, 1955; Mitchell & Singh, 1974; Wish & Carroll, 1973). The perceptual salience of the sibilance feature was reduced in several studies (Horii

This research was supported in part by Biomedical Research Support Grant RR-07042 to the University of Maryland from the Division of Research Resources, National Institutes of Health, Public Health Service. The computer time for this project was supported in part through the facilities of the Computer Science Center of the University of Maryland. The author is grateful to Tim Bunnell for development of the waveform analysis software used in this study, to Stanley Weiss for implementing the multidimensional scaling analyses, and to Peter Fitzgibbons for critiquing an earlier draft of this manuscript.

The author's mailing address is: Department of Hearing and Speech Sciences, University of Maryland, College Park, MD 20742.

et al., 1970; Miller & Nicely, 1955; Singh, 1973; Wish & Carroll, 1973), but not in that of Mitchell and Singh (1974). An acoustic interpretation of dimensions perceived in white noise has revealed that the usefulness of the features "periodicity/burst order" and "first formant transition" increases in white noise, whereas that of the features "second formant transition" and "spectral dispersion" decreases (Soli & Arabie, 1979). In addition, a significant vowel effect has been observed when the coarticulated vowel was manipulated: consonants followed by /u/ and /a/ were more frequently identified correctly in white noise than were consonants followed by /i/ (Wang & Bilger, 1973). This has been attributed to either direct masking of the high-frequency second formants in /i/ or to the less distinctive formant transitions in consonants paired with /i/.

Although both phonetic and acoustic interpretations of dimensions retrieved from performance suggest similar spectral masking effects of white noise, the basis of listener's perceptual judgments in noise is unclear. Soli and Arabie (1979) demonstrated that small consistent differences in perception of similar phonemes are evident in white noise, which supports the view that perception of phonemes in white noise is based primarily on acoustic masking effects. The constant power spectrum of white noise predictably causes perceptual effects based on acoustic properties, because different levels of white noise selectively eliminate different acoustic portions of the speech signal. Other types of interfering signals may not create exclusively acoustic masking effects. Thus, it is difficult to extend the concept that acoustic properties of speech sounds are the basis for perceptual judgments in noise to perceptual effects obtained in interference conditions other than white noise. A test of this issue would be to use a type of background that has the potential for providing interference effects other than pure acoustic masking. Multitalker babble is one type of interference that could satisfy this criterion. Unlike white noise, babble is a patterned interference which contains phonetic, linguistic, and semantic cues. Several investigations have shown that speech babble interferes with speech intelligibility more than does a continuous nonspeech noise (Carhart, Johnson, & Goodman, 1975; Carhart, Tillman, & Greetis, 1969; Speaks, Karmen, & Benitez, 1967). Kalikow, Stevens and Elliott (1977) have inferred from these findings that the greater interference of babble is partially attributed to masking specific to speech cues. Because babble has the potential for creating interference exceeding that caused by acoustic masking alone, it provides a stronger test of the notion that acoustic cues alone form the basis of observed consonant confusions in noise.

The purpose of this investigation was to determine whether perceptual relationships observed between consonant phonemes in multitalker babble backgrounds were based on a phonetic feature system or on the acoustic properties of the stimuli. Consonants were presented in three different vowel contexts and three different levels of multitalker babble to determine if certain properties of speech sounds emerged as important in a number of different listening conditions. In addition, vowel context and noise level were significant factors contributing to consonant recognition performance in white-noise conditions. Multidimensional scaling analysis was used to identify perceptual relationships among stimuli without imposing a priori any framework of relevant perceptual features. If a phonetic feature system forms the basis of perceptual judgments, then consonants should emerge in discrete clusters on each dimension. Consonants within each cluster would share a single attribute of the feature specified by the dimension, and different clusters would represent different values of the feature. Conversely, if acoustic properties of the stimuli underlie consonant perception, then the consonants should emerge either in discrete clusters or in a continuously dispersed series, depending upon the characteristic of the acoustic property associated with the dimension.

METHOD

Subjects

Listeners were selected for the experiment on the basis of age and hearing status. Ten undergraduate students from the University of Maryland participated. The subjects ranged in age from 18 to 28 years and had had no prior experience in speech-perception experiments. Each subject's pure-tone detection thresholds were assessed on a Grason-Stadler 1704 clinical diagnostic audiometer via standard audiometric techniques. The detection thresholds of all subjects were ≤ 15 dB HTL (ANSI, 1969) at octave frequencies from .25 through 8 kHz. All subjects were paid for their participation in the study.

Stimuli

The stimuli were selected to represent all individual consonant phonemes of English that occur in the initial position of words. The consonants included: /b,d,g,p,t,k,m,n,s,z,ʃ,v,f,θ,ð,r,w,j,l/. Each consonant was paired with /a/, /i/, and /u/ in a CV sequence, for a total of 57 CVs. The stimuli were spoken by a trained speaker of General American dialect and recorded onto an Otari MX5050B tape recorder in an anechoic chamber. A carrier phrase was not recorded.

Careful control of the relative levels of the signals and maskers is important in any masking experiment. The levels of all CV signals were equated in this experiment by computer adjustment to ensure equivalent S/Ns across stimuli for a given noise condition. The recorded stimuli were low-pass filtered (5-kHz cutoff, 48-dB/octave attenuation rate), digitized onto a PDP-12 laboratory computer (11.43-kHz sampling rate), and the RMS energy of each CV was determined over 20-msec intervals. All stimuli were then scaled in level so that their peak levels (integrated over 20 msec) were equivalent.[1] Because the peak energy of each CV was in the vowel portion, the effect of this procedure was to equate vowel amplitude and preserve intrinsic consonant amplitude.

The peak-equivalent digitized CVs were converted to analog signals (11.43-kHz rate), low-pass filtered (5-kHz cutoff, 48-dB/octave attenuation rate), and recorded in five different randomizations. For each tape, the 57 CVs were presented 10 times in randomized order, with a 2-sec interstimulus interval (ISI). A 1-kHz calibration tone was recorded at the beginning of each tape. This calibration tone was equivalent in RMS to the peak RMS level of each stimulus.

Masker

The interfering signal was the 12-talker babble of the SPIN test (Kalikow et al., 1977). This babble consists of six male and six female voices, and has been equated so that maximum level fluc-

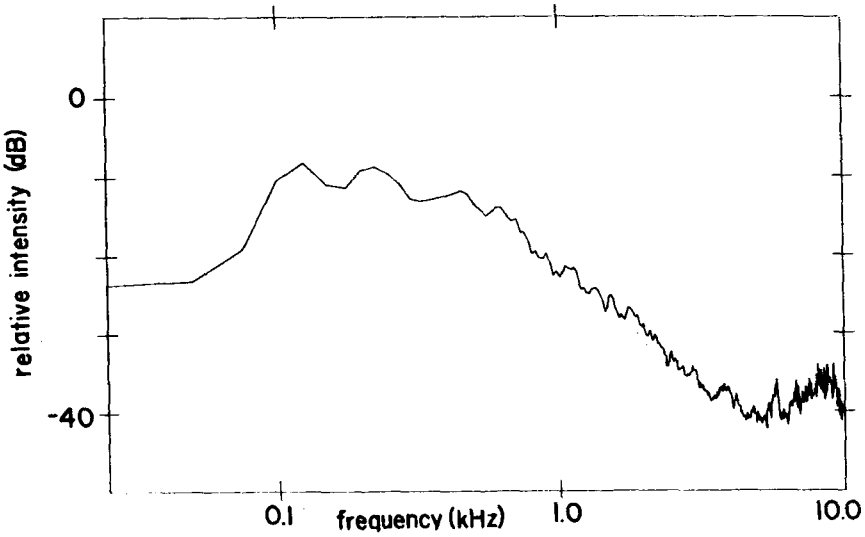

Figure 1. Long-term average spectrum of the 12-talker babble.

tuations do not exceed ±4 dB of the baseline. The long-term spectrum of the babble, as measured by a Nicolet 446AR spectrum analyzer, is shown in Figure 1.

Procedure

During each listening condition, one randomization of the 570 CVs was presented to a listener over TDH-49 earphones. Four experimental conditions were run. These included a quiet condition and three noise conditions: +12 dB signal-to-noise ratio (S/N), +6 dB S/N, and 0 dB S/N. The S/Ns were selected after pilot testing to obtain a range of performance scores. The order of listening conditions was randomized across subjects. Two practice sessions were conducted prior to the experimental sessions, to familiarize the subjects with the stimuli and procedures. In each practice session, 10 57-CV randomizations were presented to the listeners in quiet.

Calibration procedures were conducted prior to each experimental run. CVs were adjusted in level so that a 1-kHz pure tone of equivalent RMS produced 80 dB SPL at the output of the earphones. The overall level of the babble was calibrated to produce the appropriate sound pressure level at the output of the earphones, for each S/N. In noise conditions, the signals and maskers were separately calibrated, mixed, and presented monaurally.

The listener's task was to circle the initial consonant in a test booklet, which displayed the 19 consonants in alphabetical order and orthographic form. Feedback was not provided. All testing was conducted in a double-walled sound-isolated chamber. The entire test procedure was completed in 4½ h scheduled over three listening sessions.

RESULTS

General Performance

The percent-correct recognition score means and standard deviations for the 10 subjects in each listening condition are presented in Table 1. Scores obtained in the quiet condition are approximately 5%-10% lower than those reported by Wang and Bilger (1973), who used similar stimuli and response alternatives. The comparatively high scores obtained by Wang and Bilger may be attributed to the use of feedback in that experiment.

Each subject's percent-correct score in each listening condition was transformed to arcsin units to achieve homogeneity of error variance and normality of treatment-level distributions. The arcsin transform is commonly used when means and variances are proportional (Kirk, 1968). The transformed scores were submitted for analysis of variance to determine the overall effects of noise level and vowel context on consonant recognition. A significant main effect was found for noise condition [$F(3,27) = 316.65$, $p < .0001$] and for vowel [$F(2,18) = 32.12$, $p < .0001$]. No significant interactions were observed. Multiple comparison testing revealed that performance decreased significantly as noise level increased, and that identification of consonants paired with /a/ was significantly poorer than of consonants paired with /i/ or /u/.

Acoustic Analyses

One purpose of this experiment was to determine whether perception of the stimuli could be accounted for on the basis of their acoustic properties. In preparation for interpretation of the multidimensional scaling analysis (described below), acoustic analysis of the stimuli was

Table 1
Mean Percent-Correct Nonsense Syllable Recognition Scores and Standard Deviations from 10 Subjects

Vowel Context	Listening Condition: Quiet Mean	Quiet SD	S/N +12 dB Mean	S/N +12 dB SD	S/N +6 dB Mean	S/N +6 dB SD	S/N 0 dB Mean	S/N 0 dB SD
/a/	79.84	6.46	59.05	7.54	42.53	10.64	19.00	3.67
/i/	82.89	7.87	70.16	6.65	54.62	9.87	27.16	5.03
/u/	81.74	8.89	71.32	6.32	56.21	10.29	28.05	6.38

conducted. To this end, each recorded CV was low-pass filtered (7-kHz cutoff frequency) and digitized onto an LSI 11/23 microcomputer system (16-kHz sampling rate).[2] Waveform analysis programs were used to measure 22 acoustic characteristics of each CV that were considered to be potentially important to consonant perception.

The first measurement, total duration, was obtained by segmenting the onset and offset of the stimulus waveform and calculating the duration between these two boundaries. The boundary between the consonant and vowel portions was then identified from a visual display of the waveform and confirmed auditorily. Consonant duration was calculated between syllable onset and CV boundary; vowel duration was calculated between CV boundary and syllable offset.

Fundamental frequency (F0) and RMS amplitude were calculated on a 40-msec time window that stepped through the waveform file in 10-msec steps. Voicing onset was determined from these data as the time following syllable onset when the fundamental frequency was first identified. The duration of aperiodic energy between onset of the syllable and onset of the vowel was the burst duration. In the F0 analysis, the presence of initial energy at frequencies other than the fundamental marked the presence of the burst. Total consonant energy was calculated in one step by summing the log of the energy in all of the windows within the entire segment of the consonant. Vowel energy was calculated using a similar procedure over the segment of the vowel. The ratio of consonant to vowel energy was calculated as the consonant-vowel (CV) ratio.

Linear prediction coefficient (LPC) analysis (Markel & Gray, 1976) was used to compute the formant frequencies of each CV. The analysis was performed using a 25-msec Hamming window, which was stepped through the file in 10-msec steps. Formant frequencies and relative amplitudes were computed for each frame of the analysis. By viewing these data, the first and second formant (F1 and F2) transition starting frequencies and ending frequencies could be identified. F1 and F2 transition durations were calculated as the duration between the starting and ending frequencies. F1 and F2 magnitudes were defined as the change in frequency, in hertz, from the start of the transition to the end of the transition. The slopes of F1 and F2 were computed by dividing the transition magnitude by the transition duration. In addition, F1 onset was defined as the time following syllable onset when the first formant transition began. The LPC formant frequency estimates also enabled identification of the two spectral peaks of the highest amplitude present during the consonant segment of the syllable. These peaks were labeled P1 and P2. Periodicity/burst was defined as the time between the onset of voicing and the onset of aperiodic energy. Negative values signified that the onset of voicing preceded the onset of the burst; positive values indicated that the onset of the burst preceded voicing onset. These values were calculated from both the F0 analysis and the formant analysis, to enable identification of the fundamental and burst onsets.

ALSCAL Analysis Procedure

To understand the specific effects of multitalker babble on consonant perception, the pattern of errors was analyzed by multidimensional scaling techniques. The individual differences multidimensional scaling model of the ALSCAL-4 program package (Young & Lewyckyj, 1979) was used for this purpose. The individual differences scaling model has been described thoroughly by other investigators (Carroll & Chang, 1970; Walden & Montgomery, 1975; Wish & Carroll, 1973). In the present analysis, confusion matrices were prepared separately for each of the three noise conditions in each of the three vowel contexts, by pooling the confusion data of all 10 subjects. Nine matrices resulted, each representing 1,900 observations of the stimuli. These confusion matrices were log transformed to remove the discrepancies between the model's assumption of a linear relationship between estimated distances and input confusions, and the observed exponential decay function describing the relationship between derived distances and input confusions (Arabie & Soli, 1982). The matrices also were symmetrized prior to analysis, according to Shepard's (1972) formula, in which the total number of confusions between two stimuli are divided by the total number of correct responses to these same two stimuli.

The ALSCAL program analyzed the confusion matrix data to create a spatial representation of the stimulus objects, in two or more dimensions. This "group stimulus space" represented the perceptual dimensions that were common to all subjects in all noise conditions in perceiving the stimuli. ALSCAL also constructed a "condition space," in which the noise conditions × vowel contexts are represented as vectors of weights in a second multidimensional space. These dimension weights indicate the importance of the various dimensions of the group stimulus space, for each condition. The ALSCAL-4 program uses an iterative, alternating least squares procedure to determine the stimulus coordinates and dimension weights that account for the maximum possible variance in the confusion matrix data.

Solutions were obtained in two through six dimensions. The group stimulus space revealed interpretable dimensions for the five-dimensional solution. In addition, the solution in five dimensions accounted for greater variance than solutions in fewer dimensions, and approximately the same amount of variance as the six-dimensional solution. The five-dimensional solution accounted for 63.2% of the variance in the confusion data, indicating that the solution fit the data reasonably well.

Stimulus Configuration

The stimulus configuration for the five-dimensional solution is depicted in Figure 2. This configuration represents the perceptual weightings of the stimuli that were common to all subjects in the three noise conditions across

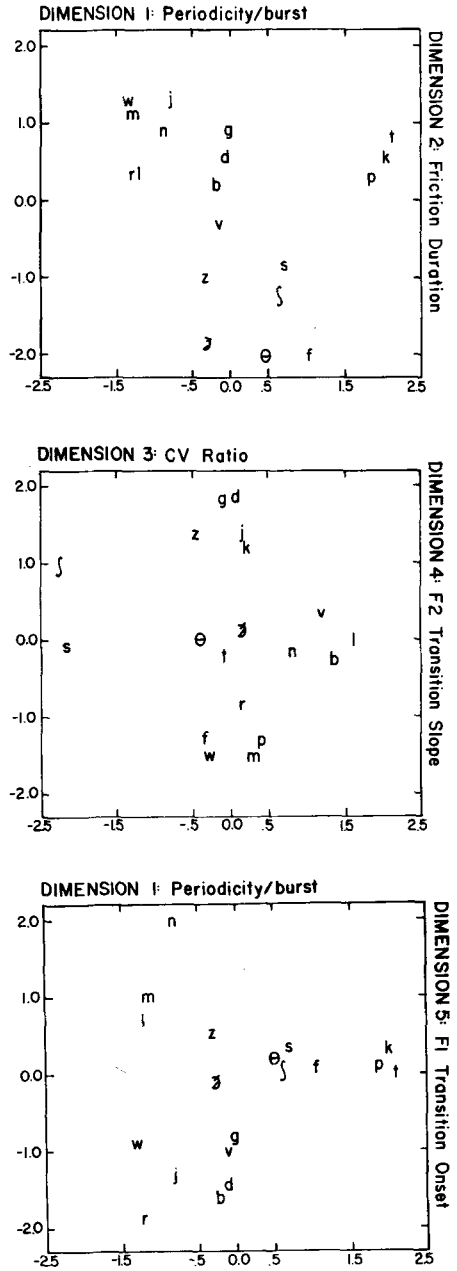

Figure 2. Five-dimensional group stimulus configuration derived by ALSCAL.

the three vowel contexts. Two approaches will be used to interpret each dimension.

The first approach involves identification of a phonetic feature that might account for the arrangement of stimuli on the dimension. If the dimension can be accounted for on the basis of a distinctive phonetic feature, then the stimuli should appear in discrete clusters corresponding to the binary or tertiary values of that feature.

The second approach is to identify an acoustic property which best corresponds to the observed relations among the stimuli. To that end, Pearson product-moment correlation coefficients were computed between each stimulus for each dimension and each of the 22 acoustic values previously determined for each CV. Intercorrelations among these variables also were calculated. The derived correlation matrix was used as input to a least squares multiple linear regression analysis. The purpose of the multiple regression procedure was to find an equation that identified the acoustic property or properties that contribute most to the stimulus coordinates, while controlling for other confounding factors. Independent variables were entered into the analysis if the resulting equation produced an F-ratio whose probability was less than .05. Variables that did not improve R^2 were restricted from the equation. Table 2 shows the acoustic properties in the equations which predicted the stimulus weights. The partial correlation coefficients of each property and the associated R^2 for each equation are shown also. The level of statistical significance achieved for each of these prediction equations was $< .01$. For each dimension, only one acoustic property contributed significantly to the prediction equation across all three sets of consonants. These acoustic features were selected as the acoustic property underlying the associated dimension.

The plot of stimuli on Dimension 1 (D1) roughly corresponds to the phonetic feature voicing. Stimuli appearing on the right half of the space are voiceless; stimuli appearing on the left half of the space are voiced. However, the voicing feature is insufficient to explain the further separation of the voiceless stimuli /t,k,p/ from the voiceless stimuli /f,s,ʃ,θ/ and the separation of the voiced stimuli /g,d,b,v,z,ð/ from the voiced stimuli /j,n,l,r,w,m/.

The acoustic feature that is most highly correlated with the stimulus weights on D1 is periodicity/burst. High positive correlation coefficients between acoustic values and stimulus weights indicated that stimuli projecting to the far right of the space have bursts preceding the onset of periodicity, whereas stimuli projecting to the left of the space have a period of prevoicing (i.e., periodicity precedes the burst onset). The nasals and sonorants, which are characterized by prevoicing, are plotted appropriately on the left of the space. Voiceless plosives project toward the right of the space because the onset of the burst precedes the onset of periodicity by as much as 50 msec in these consonants (Lisker & Abramson, 1964). The voiced plosives and fricatives appear in the center of this continuum, because the onsets of the burst and periodicity are nearly simultaneous among these stimuli (Lisker & Abramson, 1964). The voiceless fricatives are plotted near the voiceless plosives, which is consistent with the onset of friction noise preceding the onset of formant resonance in these stimuli. However, one might expect the voiceless fricatives to be displayed closer to the boundary of the continuum than the voiceless plosives, because the temporal interval between the two acoustic events is greater among the voiceless fricatives than the voiceless plosives. The rapidity of burst onset is a second acoustic property which may account for this result. Specifically, the rapid onset of the burst in the voiceless plosives may

Table 2
Partial Correlation Coefficients and R^2 Values of Selected Stimulus Acoustic Properties That Predict Stimulus Dimension Weights, Identified from Multiple Regression Analysis

	/a/			/i/			/u/		
Dimension	Acoustic Property	Partial r	R^2	Acoustic Property	Partial r	R^2	Acoustic Property	Partial r	R^2
1	periodicity/burst	.84	.82	periodicity/burst	.93	.90	periodicity/burst	.87	.80
	total duration	−.74		F2 transition duration	−.84		F1 onset	−.72	
	F1 magnitude	−.63		F1 transition duration	.76				
				F2 magnitude	−.78				
2	friction duration	−.88	.77	friction duration	−.88	.85	friction duration	−.84	.80
				vowel energy	−.76		F1 transition duration	.62	
3	CV ratio	−.76	.58	CV ratio	−.88	.80	CV ratio	−.84	.80
				voicing onset	−.61		friction duration	.69	
				vowel duration	.49				
4	F2 slope	−.82	.71	F2 slope	−.64	.50	F2 slope	−.58	.34
	F2 starting frequency	.75		F1 slope	.58				
5	F1 onset	.63	.39	F1 onset	.74	.57	F1 onset	.68	.53
				friction duration	−.60		F1 ending frequency	−.55	

have served as a secondary perceptual cue in distinguishing the voiceless stimuli from the nasals and sonorants. The gradual onset of friction noise in the voiceless fricatives may have reduced the prominence of the temporal order cue of these phonemes. D1 was labeled "burst/periodicity order" in accordance with the term used by Soli and Arabie (1979) to describe a similar dimension.

A phonetic feature that could account for the stimulus configuration for D2 is frication. All the fricatives among this stimulus set /f,θ,ð,ʃ,z,s,v/ have negative coordinates on D2; all nonfricatives have positive coordinates. The limitation of this phonetically based interpretation is that the arrangement of the fricatives is diffuse. That is, distinctions are apparent between three subsets of fricatives: /θ,ð,f/, /z,s,ʃ/, and /v/, which could not be explained by a phonetic account.

Alternatively, the acoustic feature friction duration correlates well with the stimulus weights for D2. This label refers to the duration of aperiodic energy at the onset of the syllable. High negative correlation coefficients were observed, indicating that stimuli with low weights on D2 have longer durations of aperiodic energy than stimuli with higher weights on the dimension. The fricatives and sibilants among this stimulus set contain aperiodic energy with durations of 150 to 200 msec. The exception to this observation is /v/, which has a duration of friction noise of only 30 msec. Consequently, it is distinguished from the other fricatives on the dimension. Projection of the stimuli /w,j,m,n/ at the top of the dimension is consistent with the absence of aperiodic energy among these consonants. However, the presence of the voiceless plosives /p,t,k/ near the top of the dimension is somewhat incongruent with this interpretation, because they contain a burst of aperiodic energy of approximately 50 msec duration. One explanation may be that the relatively brief bursts among these consonants were partially obscured in the various noise conditions.

Projection of the stimuli on D3 shows a clear separation of the sibilants/ʃ,s/ from all other stimuli. Although the phonetic feature sibilance is an obvious label for D3, it does not account adequately for the appearance of /z/ among the nonsibilant stimuli.

The regression analysis revealed high negative correlation coefficients between the stimulus weights and their CV ratios for D3. The sibilants are characterized by strong aperiodic energy of long duration, resulting in comparatively high CV ratios. Phonemes projecting high on the dimension, /v,n,b,l/, have CV ratios that are approximately 20 to 25 dB lower than those of the sibilants. Stimuli that appear toward the middle of the dimension generally have CV ratios that are intermediate values, but which are still 12 to 17 dB below those of the sibilants. Thus, it appears that distances between stimuli on this dimension correspond closely to differences between CV ratios among them.

Three clusters of stimuli appear on D4. The cluster near the bottom of the space, /f,w,p,m,r/, contains stimuli produced with maximum constriction at the lips, with the exception of /r/. The stimuli comprising the middle cluster, /s,θ,ð,t,n,v,b,l/, are produced with maximum constriction at the alveolar ridge or interdentally, with the exception of /v,b/. The cluster at the top of the space contains stimuli produced with maximum constriction at the back of the mouth, /g,k,j,ʃ/, as well as stimuli produced at the alveolar ridge, /z,d/. Thus, a phonetic feature that could be assigned to this dimension is place. However, it is clear that the projection of stimuli within the three place clusters is imprecise.

The acoustic property that correlates best with the stimulus weights on D4 is F2 transition slope. We note that consonants present in the top cluster are characterized by falling F2 transitions, whereas consonants present in the bottom cluster have rising F2 transitions. The stimuli plotted in the middle cluster have relatively flat or shallow F2 transitions. A view of the condition space indicates that D4 was weighted much higher when consonants were paired with /a/ than when they were paired with /i/ or /u/. Thus, the ordering of phonemes on D4 pertains

primarily to the consonants produced in the /a/ vowel context. Acoustic analysis of consonants paired with /a/ indicates that the slope of the F2 transition closely parallels the arrangement of stimuli on D4. Specifically, the stimuli /pa,ma,wa,ra,fa/ exhibit rising F2 transitions; the stimuli /ba,ta,na,la,sa,va,ða,θa/ exhibit relatively straight F2 transitions; and the stimuli /da,ga,ka,ja,ʃa,za/ display falling F2 transitions.

Examination of the arrangement of stimuli on D5 does not readily suggest a corresponding phonetic feature. Although nasality could be considered because of the prominence of /n/, the separation of /n/ and /m/ precludes this interpretation. Furthermore, there are two additional clusters of stimuli projecting in the middle and bottom of the space, which could not be accounted for by the nasality feature.

Stimulus coordinates for D5 are correlated most highly with the acoustic characteristic F1 onset. Stimuli plotted near the bottom of the space, /b,d,g,v,w,r,j/, have little or no delay in the onset of the first formant. Stimuli appearing near the middle of the space include the voiceless plosives /p,t,k/, the voiceless fricatives /f,θ,s,ʃ/, and the voiced fricatives /z,ð/. The voiceless stimuli are characterized by a burst of spectrally dispersed energy with an associated delay in the onset of the first formant (Lisker, 1975; Stevens & Klatt, 1974). The voiced fricatives in this stimulus set also exhibit a delay in F1 onset. Stimuli appearing near the top of the space include the nasals /m,n/ and the sonorant /l/. These stimuli are produced with a period of prevoicing, during which there is no apparent formant structure. The nasals also contain a period of nasal resonance, due to the side-branching of the nasal passage from the vocal tract. During the period of nasal resonance, the nasals do not exhibit rapid spectral change. Thus, the presence of prevoicing and/or nasal murmur delays the onset of a well-defined F1 transitional period.

Condition Space

Although the stimulus configuration represents the common dimensions used by the subjects in the three noise conditions × three vowel contexts, these dimensions may not be equally important at each listening condition. The output of ALSCAL provides a plot of the relative weightings of each dimension at each listening condition. These results are plotted in Figure 3. The effect of noise level and vowel context at each dimension can be discerned from these data.

The plot of condition weights on D1 reveals that the perceptual importance of the periodicity/burst cue increases in severe noise conditions for consonants paired with /i/ and /u/. The initial burst of noise following articulatory release is composed of spectrally dispersed energy. High-frequency energy in the burst of certain consonants is a part of this cue that is not masked by the babble's low-frequency spectrum. Thus, detection of the onset of the initial burst of noise appears to be a sufficient cue to account for the observed results. The condition space also shows that the weights for D1 decrease as noise level increases for consonants followed by /a/. Spectral analysis of the voiceless plosives revealed that voiceless plosives coarticulated with /a/ have a weaker initial burst than those coarticulated with /i/ and /u/. Because of these comparatively weak bursts, the onset of the burst is not distinguished well in severe noise conditions. The result is a decrease in the perceptual usefulness of the burst/periodicity feature as the level of multitalker babble increases.

The perceptual importance of friction duration (D2) is constant in all multitalker babble conditions, as evidenced by high weights in the /i/ and /a/ contexts and low weights in the /u/ context. This cue is sufficiently robust for increases in background noise not to change its usefulness to listeners. The stability of the friction cue across noise level may be attributed to the quasi-random nature of the friction noise. Specifically, the quasi-random noise in the fricatives may be perceptually distinct from the overall quasi-periodic components of the speech babble. A constant vowel effect is observed in the condition weights: consonants paired with /u/ have lower weights than consonants paired with /a/ or /i/. Fricatives paired with /i/ and /a/ have a concentration of energy at higher frequencies than do fricatives paired with /u/. Thus, the babble has an increasing masking effect as the friction cue is lowered in frequency.

The weights of the CV ratio cue (D3) increase in all

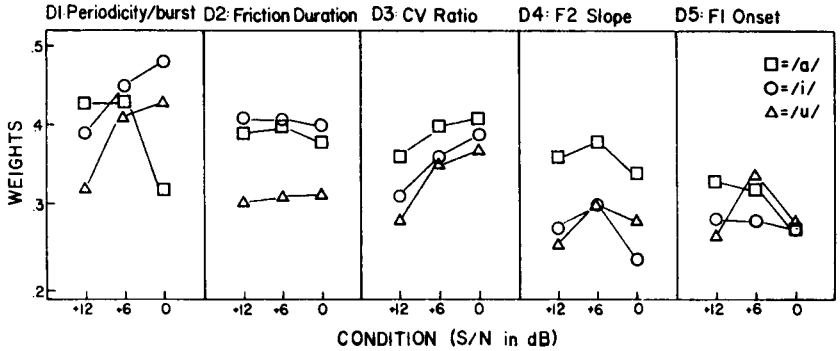

Figure 3. Weights of three noise conditions and three vowel contexts for five dimensions derived by ALSCAL.

three vowel contexts as noise level increases. Stimuli with high CV ratios are subjectively louder than those with lower CV ratios. In noise conditions, the louder consonants continue to be heard and recognized with a high degree of accuracy, whereas softer consonants become less perceptible. The babble therefore serves to increase the distinction between stimuli with high CV ratios and stimuli with low CV ratios.

The condition space shows that the F2 transition slope cue (D4) was weighted highest when consonants were produced with /a/, in all noise conditions. This result can be explained by an examination of the effect of vowel coarticulation on second formant transitions. The second formant of /u/ is lower in frequency than that of /a/ or /i/ (Peterson & Barney, 1952). The F2 transitions from consonant to vowel are primarily straight or falling transitions in the /u/ context. Conversely, the /i/ vowel has the highest second formant of all the vowels (Peterson & Barney, 1952). Rising or straight transitions occur when /i/ follows a consonant phoneme. The absence of rising F2 transitions among the /u/ stimulus set and the absence of falling F2 transitions among the /i/ stimulus set may have minimized the availability of the F2 slope cue in these contexts. The vowel /a/ has a midfrequency F2. Identifiable rising and falling F2 transitions occur when /a/ follows a consonant (Liberman, Delattre, Gerstman, & Cooper, 1956). The condition weights suggest that the F2 transitions in the /a/ stimulus set did provide distinguishing information for place. However, the weights for this cue decreased slightly as noise level increased, indicating that the relevant midfrequency acoustic information was masked in extreme noise conditions.

A perceptual effect of vowel and noise is also evident in the pattern of weights for D5. The weights for the first formant onset cue are greater for consonants followed by /a/ than for consonants followed by /i/ and /u/, in low levels of babble. The first formant of /a/ is approximately 730 Hz, which is higher in frequency than the first formants of /i/ and /u/ (Peterson & Barney, 1952). The peak amplitude of the babble is between 400 and 500 Hz, as revealed by spectral analysis (see Figure 1). Consequently, the first formant of consonants followed by /a/ is not masked extensively by low levels of babble. In the most degraded noise condition, the intensity of the babble is high enough to have a direct masking effect on the first formant in consonants followed by /a/. This is reflected in the condition space as a decrease in the weights of D5 for consonants paired with /a/. The first formants of /i/ and /u/ are sufficiently low in frequency to be masked extensively in all levels of babble. Thus, the condition weights for these stimulus sets remain stable as the level of babble increases.

DISCUSSION

The present study demonstrates that perception of consonant phonemes in multitalker babble is more adequately represented by an acoustic cue analysis than by a phonetic feature analysis. One aspect of the data that supports this notion is that, for each dimension, stimuli did not project in discrete clusters corresponding to the separate categories that comprise a distinctive phonetic feature. Additional evidence for an acoustic feature interpretation is that for all dimensions, acoustic cues were identified that exhibited moderate or strong correlations with the stimulus coordinates. Although more than one acoustic cue could have been identified as contributing to the perceptual pattern for one stimulus set, only one acoustic cue predominated across all three stimulus sets for all dimensions. These results suggest that small acoustic differences between stimuli were consistently available to listeners in the babble.

Interpretation of performance based on acoustic properties of consonant phonemes rather than distinctive phonetic features corresponds with a previous report of perceptual effects in a white-noise background (Soli & Arabie, 1979). Thus, regardless of whether the interference has a constant power spectrum or multiple and fluctuating speech cues, recognition of consonants in such degraded environments does not appear to depend on a perceptual strategy based exclusively on distinctive phonetic features. Indeed, the correspondence between acoustic parameters and consonant confusions observed in contrasting interference conditions implies that the low-level acoustic properties of the stimuli are the relevant basis for listeners' judgments. One unexpected finding of the present investigation is that the predominant acoustic cues underlying perception are comparable to those identified previously in white-noise conditions (Soli & Arabie, 1979). However, these specific attributes do not exhaust the list of potentially important acoustic cues that may be revealed under other types of degraded listening conditions.

Despite the identification of comparable acoustic cues of importance in white noise and multitalker babble backgrounds, the interference effects of the different noises are generally contrastive. Most studies that have evaluated consonant feature perception in white noise used a single vowel context. Because vowel coarticulation exerted strong effects on the perceptual importance of each dimension, direct comparisons will be made for consonants presented in the same vowel context. The studies of Soli and Arabie (1979) and Wish and Carroll (1973) were reanalyses of the Miller and Nicely (1955) data, in which consonants followed by /a/ were presented in white noise. The periodicity/burst and first formant dimensions of Soli and Arabie (1979) and the comparable voicing and nasality dimensions of Wish and Carroll (1973) retained their usefulness under severe degradation by white noise. In contrast, the importance of the periodicity/burst and F1 onset dimensions decreased for consonants followed by /a/ in conditions degraded by babble. In white-noise backgrounds, the F2 transition and spectral dispersion (sibilance) dimensions decreased in perceptual salience (Soli & Arabie, 1979; Wish & Carroll, 1973). The importance of comparable dimensions retrieved in multitalker bab-

ble increased (CV ratio), remained stable (friction duration), or decreased slightly (F2 slope) in severe levels of babble. It is interesting to note that the F2 transition cue decreased in importance in both multitalker babble and white noise, even though the spectral interferences of the two types of noise are generally contrastive. The F2 transition cue spans low, middle, and high frequencies, depending upon the particular place of vocal tract constriction. Thus, in either type of interference, some of this acoustic information will be masked.

The vowel effect in multitalker babble is different from that observed in white noise. Wang and Bilger (1973) reported that in white noise, recognition of consonants followed by /i/ was poorer than recognition of consonants followed by /a/ or /u/. The white noise produced high-frequency spectral interference on the high-frequency second formant transitions of consonants coarticulated with /i/. The current results revealed that coarticulation with /a/ yielded the poorest percentage correct recognition performance. Furthermore, the ALSCAL analysis demonstrated that coarticulation with different vowels differentially affected the perceptual importance of each dimension. For example, coarticulation with /u/ resulted in consistently low weights on the friction duration, F2 slope, and CV ratio dimensions. Low-frequency vowel formants of /u/ lowered the consonant's burst of transition cues. As a result, the acoustic cues became masked when the low-frequency spectrum level of the babble approached that of the speech stimulus. In contrast, the importance of the F2 slope cue was comparatively high for consonants coarticulated with /a/. As noted earlier, consonants coarticulated with /a/ have well-defined rising and falling F2 transitions, unlike consonants coarticulated with /i/ or /u/. Weights were notably high for consonants produced with /i/ for the periodicity/burst dimension. The presence of babble did not affect the availability of this cue because the high-frequency noise bursts in consonants followed by /i/ were above the babble's low-frequency spectrum. Generalizations about noise effects on perception of important consonant features must be limited to the specific vowel environment in which the effects are observed.

To summarize, patterns of consonant phoneme perception in a multitalker babble background demonstrate that listeners perceive and use small acoustic differences between phonemes in consonant recognition. This finding is consistent with an interpretation of observed perceptual relationships of consonants obtained in white-noise conditions. These comparable results, observed under widely contrasting conditions of degradation, suggest that acoustic representations of speech sounds more adequately account for listeners' confusions in noise than do phonetic representations. Earlier in this report it was stated that a single phonetic-feature-based system has not emerged to account for consonant-recognition performance in noise among earlier independent research studies. Although similar acoustic properties emerged in the present analysis and that of Soli & Arabie (1979), this does not necessarily imply that an invariant set of acoustic features exists to account for speech perception in noise. Further evidence is needed to determine whether comparable acoustic properties emerge when speech sounds are produced by different talkers and presented in different vowel environments, syllable sequences, and noise conditions. Despite similar encoding strategies used by listeners in multitalker babble and white noise, the interference pattern in multitalker babble is clearly different from that observed in white noise. In multitalker babble, perception of low-frequency cues of consonant phonemes is degraded, suggesting that the low-frequency long-term spectrum of multitalker babble is responsible for its masking effects. However, dramatic alterations in the perceptual importance of consonant features in noise occur with changes in vowel context. This is a result of the multiple effects of vowel coarticulation on the acoustic properties of consonants.

REFERENCES

ANSI (1969). *Standard specifications for audiometers* (ANSI S3.6-1969, R-1973). New York: American National Standards Institute.

ARABIE, P., & SOLI, S. (1982). The interface between the types of regression and methods of collecting proximity data. In R. G. Golledge & J. N. Rayner (Eds.), *Proximity and preference: Problems in the multidimensional analysis of large data sets,* (pp. 90-115). Minneapolis: University of Minnesota Press.

CARHART, R., JOHNSON, C., & GOODMAN, J. (1975). Perceptual masking of speakers by combinations of talkers. *Journal of the Acoustical Society of America*, **58**, 35A.

CARHART, R., TILLMAN, T. W., & GREETIS, E. (1969). Perceptual masking in multiple speech backgrounds. *Journal of the Acoustical Society of America*, **45**, 694-703.

CARROLL, J. D., & CHANG, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Pscyhometrika*, **35**, 238-319.

HORII, Y., HOUSE, A. S., & HUGHES, G. W. (1970). A masking noise with speech-envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America*, **49**, 1849-1856.

KALIKOW, D. N., STEVENS, K. N., & ELLIOTT, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, **61**, 1337-1351.

KIRK, R. E. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Brooks/Cole.

LIBERMAN, A. M., DELATTRE, P. C., GERSTMAN, L. J., & COOPER, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, **52**, 127-137.

LISKER, L. (1975). Is it VOT or a first formant transition detector. *Journal of the Acoustical Society of America*, **57**, 1547-1551.

LISKER, L., & ABRAMSON, A. S. (1964). A cross-language study of voicing in initial stops: Acoustic measurements. *Word*, **20**, 384-422.

MARKEL, J. D., & GRAY, A. H., JR. (1976). *Linear prediction of speech.* Berlin: Springer.

MILLER, G. A., & NICELY, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338-352.

MITCHELL, L. M., & SINGH, S. (1974). Perceptual structure of 16 prevocalic English consonants sententially embedded. *Journal of the Acoustical Society of America*, **55**, 1355-1357.

PETERSON, G. E., & BARNEY, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

SHEPHARD, R. N. (1972). Psychological representation of speech sounds. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 67-113). New York: McGraw-Hill.

SINGH, S. (1971). Perceptual similarities and minimal phonemic differences. *Journal of Speech and Hearing Research*, **14**, 113-124.

SINGH, S. (1973). *A unified theory of speech perception.* Paper presented at the Annual Convention of the American Speech and Hearing Association, Detroit, MI.

SOLI, S., & ARABIE, P. (1979). Auditory versus phonetic accounts of observed confusions between consonants. *Journal of the Acoustical Society of America*, **66**, 46-59.

SPEAKS, C., KARMEN, J. L., & BENITEZ, L. (1967). Effect of a competing message on synthetic sentence identification. *Journal of Speech and Hearing Research*, **10**, 390-395.

STEVENS, K. N., & KLATT, D. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, **55**, 653-659.

WALDEN, B. E., & MONTGOMERY, A. A. (1975). Dimensions of consonant perception in normal and hearing-impaired listeners. *Journal of Speech and Hearing Research*, **18**, 444-455.

WANG, M. D., & BILGER, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America*, **54**, 1248-1266.

WISH, M., & CARROLL, J. D. (1973). Applications of 'INDSCAL' to studies of human perception and judgment. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (4th ed.) (pp. 449-489). New York: Academic Press.

YOUNG, F. W., & LEWYCKYJ, R. (1979). *ALSCAL-4 user's guide.* Carrboro, NC: Data Analysis and Theory Associates.

NOTES

1. One common practice for equating stimulus level is to produce stimuli that peak at the same level on a VU meter. The method used in this experiment represents a computerized technique for modifying CVs so that they peak at equivalent levels. This technique is comparable to that employed by Wang and Bilger (1973). However, the amplitude of CVs in the /a/ context may have been reduced relative to their amplitude in natural speech with this approach.

2. Despite the fact that acoustic analysis was performed with 7-kHz low-pass filtering, the speech stimuli had a high-frequency cutoff of 5 kHz.

(Manuscript received December 30, 1983;
revision accepted for publication July 22, 1985.)