Original Paper

Phonetica 1994; 51: 195–209

Keith Johnson[a]
James V. Ralston[b]

[a] Department of Linguistics, Ohio State University, Columbus, Ohio;
[b] Department of Psychology, Ithaca College, Ithaca, N.Y., USA

Automaticity in Speech Perception: Some Speech/Nonspeech Comparisons

Abstract

Three experiments used sine-wave replicas of speech sounds to explore some differences between speech perception and general auditory perception. The experiments compared patterns of behavior in categorization and discrimination tasks for listeners reporting either speech or nonspeech percepts of sine-wave replicas of speech. We hypothesized that the perception of speech sounds is automatized, while the perception of less familiar sounds is not. The first experiment was designed to investigate the perception of relatively long initial consonant transitions using a synthetic /wa/–/ya/ sine-wave analog continuum. Speech listeners perceived the continuum categorically, but nonspeech listeners could not consistently categorize the items in the continuum. In the second experiment, both speech and nonspeech listeners could consistently categorize stimuli having final glides (an /ay/–/aw/ sine-wave replica continuum), but differences between speech and nonspeech listeners were found in the slopes of the identification functions, in reaction times, and in the effect of context. These differences are consistent with the hypothesis that speech perception is automatized. In the third experiment, nonspeech listeners' discrimination sensitivity was greater than speech listeners'. The observed pattern of results suggests that speech perception is accomplished by a fast, obligatory, and thus automatic perceptual mechanism.

Received:
May 14, 1993
Accepted:
January 10, 1994

Keith Johnson
Department of Linguistics
222 Oxley Hall, 1712 Neil Avenue
Ohio State University
Columbus, OH 43210-1298 (USA)

© 1994
S. Karger AG, Basel

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

Shiffrin and Schneider [1977] showed that automatic perceptual processing results from long-term learning. They and others characterize automatic processing as highly efficient, obligatory, requiring little or no attention, difficult to modify, and resistant to disruption (p. 160). Subjects developed automatic processing in a visual search task when there was extensive training and stimuli were consistently mapped to responses (i.e. the same stimulus never served as both a target and a distractor). Subjects given the same degree of extensive practice under conditions in which the stimuli were not consistently mapped to responses showed controlled processing. Controlled processing requires attention (and thus is capacity-limited), is slow, serial in nature, and can be modified easily (p. 160).

Shiffrin and Schneider [1977, p. 153] also informally noted that automatic processing is independent of stimulus variability. Subjects in the consistent-mapping conditions reported that, even though the stimuli used in the training sessions were composed of upper-case characters, the targets tended to 'pop out' of ordinary text outside of the laboratory. Flowers et al. [1981] noted a similar phenomenon in an identification task involving random strings of letters. If the string contained a word, search for a random target string was inhibited. Subjects reported that the words 'popped out' of the display.

In this article we present the results of a test of the hypothesis that speech perception is an automatic process.[1] *A priori* this seems to be a reasonable hypothesis because (1) listeners have extensive experience with speech sounds, (2) speech perception is accomplished very quickly and (3) with little apparent effort, and (4) speech perception processes are difficult to modify [for instance, in second-language acquisition, Brière, 1966]. However, despite the face validity of the automaticity hypothesis it has never been tested. This fact alone is reason enough to perform the research described here. It is also important to note that several aspects of speech perception that have led to claims for the 'specialness' of speech may be attributable to automaticity (we will return to this in the 'General Discussion').

The experiments reported here used replicas of natural speech that were produced by combining sine waves which obey the regularities of frequency range, amplitude and trajectories of vocal tract resonances [Remez et al., 1981]. These speech replicas did not contain the acoustic correlates of vocal cord vibration (fundamental frequency and glottal spectrum), nor did they have the formant bandwidths typical of natural speech, and so were, in these ways, not speech-like. Yet, the preservation of speech-like formant frequencies and amplitudes, coupled with plausible formant trajectories, rendered the signals speech-like.

Previous research has shown that listeners' perceptual experiences of sine-wave replicas of speech reflect the ambiguity of the acoustic signal. These signals are not immediately recognizable as speech, but if listeners are biased to expect to hear speech, many are able to correctly transcribe the intended utterance [Remez et al., 1981]. In addition, many listeners when required to categorize sine-wave replicas of speech will spontaneously hear them in terms of speech categories [Bailey et al., 1977], although the signal's unnatural quality remains apparent. Thus, sine-wave replicas of speech provide a control condition for studying differences between perceiving an acoustic signal as speech and perceiving that same signal as a nonspeech auditory event.

[1] Although the examples of automatic perceptual processing cited earlier involve perceptual learning, we are not claiming or assuming that automaticity is necessarily learned. It would be interesting if it could be demonstrated that automaticity in speech perception (if it exists) arises without learning.

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

There have seen several recent studies comparing speech and nonspeech perception using sine-wave replicas of speech. These studies have found that speech listeners use trading relations among acoustic cues for phonetic categories while nonspeech listeners do not [Best et al., 1981]; that speech listeners are able to use acoustic information which seems to be masked for nonspeech listeners [Grunke and Pisoni, 1982; Schwab, 1981], and that speech listeners seem to integrate acoustic information over CV syllable stimuli while nonspeech listeners do not [Tomiak et al., 1987].

We made three assumptions about automatic processing and from these devised tests of the automaticity hypothesis.

First, we assumed that automatic processes are generally faster than nonautomatic processes. For instance, Shiffrin and Schneider [1977] found that subjects could locate visual targets in a search paradigm more quickly after extensive training in consistent-mapping conditions than before training or in variable-mapping conditions. This was taken as supporting evidence for their claim that training in the consistent-mapping condition produced automaticity. Shiffrin [1987] warns against relying exclusively on reaction time data in testing for automatic processing because subjects' familiarity with a task or set of stimuli may affect performance whether processing is automatic or not. In the present experiments, reaction time is a meaningful measure of automaticity because our speech and nonspeech listeners were performing identical tasks with identical stimuli.

Second, we assumed that automatic processing causes attention to be switched from sensory representations to categorical representations. This assumption leads us to predict two differences between speech listeners and nonspeech listeners: (1) If speech perception is automatic, speech and nonspeech listeners should show different contrast effects in labeling performance. Successive stimuli in an identification experiment may contrast at two levels: sensory and categorical [Pisoni, 1973]. If automatic processing produces an obligatory attentional shift to the categorical level, the basis for stimulus-to-stimulus contrast (the amount of information held over from trial to trial) will be larger in the case of nonautomatic processing because in nonautomatic processing attention may be more broadly focussed to include both sensory and categorical information. So, if speech perception is automatic we predict that nonspeech listeners will show larger contrast effects than will speech listeners. (2) Assuming that automatic processing causes attention to be switched from sensory representations to categorical representations also leads to the prediction that nonspeech listeners will out-perform speech listeners in within-category discriminations because speech listeners will not attend to the auditory/sensory representation of the signal after attention has been automatically switched to the categorical level.

Third, we assumed that if speech perception is an automatic process attention will be automatically drawn to speech-relevant acoustic dimensions of the signal, and therefore the nonspeech listener's attention will not be so narrowly or efficiently focussed as the speech listener's. This assumption leads to the prediction that speech listeners will be able to use acoustic information that nonspeech listeners cannot because the speech listener's attention is automatically drawn to phonetically relevant acoustic attributes while the nonspeech listener does not know what to focus on in a psychoacoustically complex signal like sine-wave analogs of speech. Note that this prediction seems to be at odds with our earlier prediction that nonspeech listeners' within-category discrimination performance will be better than speech listeners'. The predictions can be integrated with each other as follows: non-

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

speech listeners' within-category discrimination performance will be better than speech listeners', except when nonspeech listeners are unable to focus attention on the relevant part of the signal.

Not enough is known about the psychoacoustics of multitone, time-varying signals to be able to predict what acoustic aspects of sine-wave replicas of speech will prove to be problematic to naive nonspeech listeners. Schwab [1981] and Grunke and Pisoni [1982] found that nonspeech listeners could not label CV (/da/ /ga/) stimuli as well as they could VC (/ad/ /ag/) stimuli, an effect that Schwab [1981] likened to backward masking. On our view, the nonspeech listener's difficulty with CVs as compared with VCs suggests that the nonspeech listener is unable to focus attention on the formant transitions in CV stimuli because the following steady-state vowel proves to be distracting. This type of account also seems to fit the other two types of stimuli that have proven to be easier for speech listeners than for nonspeech listeners. Schwab [1981] found that when F_1 and F_2 change in the same direction nonspeech listeners' labeling performance is as accurate as speech listeners', but when F_1 changes in one direction while F_2 changes in the other nonspeech listeners' performance deteriorates dramatically while speech listeners are not affected. This seems to be a clear case of a failure to focus attention narrowly enough or quickly enough to perform a perceptual task. Best et al. [1989] found that speech listeners could label the items in a /ra/–/la/ sine-wave analog continuum much more consistently than could nonspeech listeners. This finding reflects not only the speech listeners' advantage for CV syllables but also for processing higher frequency components of a multitone complex (the distinction between /r/ and /l/ was signaled by F_3, while F_1 and F_2 were unchanged across the continuum). Schwab [1981] likened the nonspeech listener's difficulty with higher components (but not with the lowest of a multitone sequence) to the upward spread of masking. In our view, the effect (which clearly is not peripheral masking) is due to the nonspeech listener's fragile attention, as compared with the speech listener's well-focussed attention.

So there is abundant evidence in the literature that in some circumstances speech listeners can out-perform nonspeech listeners in the identification and discrimination of sine-wave replicas of speech. The automaticity hypothesis is interesting in that it predicts that the opposite should also occur in other circumstances.

To test these predictions we compared the perceptual responses of listeners who heard sine-wave replicas of speech as speech sounds (speech listeners) with the responses of other listeners who heard those same stimuli as nonspeech noises (nonspeech listeners). Experiment 1 extends the work of Grunke and Pisoni [1982], Schwab [1981], and Best et al. [1989] to a new continuum using an experimental method first reported by Bailey et al. [1977]. The experiment was designed to explore the speech listener's advantage with CV stimuli by using a glide-vowel continuum in which the relevant acoustic cue (the F_2 transition) could be more psychoacoustically salient than in the stop-vowel continua used in previous research. Experiment 2 explores the differences between speech and nonspeech identification performance in VC syllables and tests the predictions for reaction time and context effects. Experiment 3 tests the prediction that nonspeech listeners will show better within-category discrimination performance.

Experiment 1

One of the most striking differences between speech and nonspeech perception has

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

to do with the perception of CV syllables [Grunke and Pisoni, 1982; Schwab, 1981; Best et al., 1989]. When listeners hear sine-wave replicas of stop-vowel syllables as nonspeech they are virtually unable to use the initial F_2 or F_3 transition in a categorization task. For instance, Grunke and Pisoni [1982] presented sine-wave replicas of /ba/ and /da/ to listeners in a categorization task. They found that listeners who were biased to hear the stimuli as speech were able to consistently classify the stimuli, while listeners who were biased to hear the stimuli as nonspeech sounds responded randomly. Best et al. [1989] also found that nonspeech listeners unlike speech listeners were virtually unable to classify glide-vowel syllables which differed in terms of their initial F_3 transitions (a /ra/–/la/ continuum).

We extended these studies by investigating the perception of a /wa/–/ya/ continuum of sine-wave replicas. Because the stimuli in this continuum had longer transitions than the stop consonants employed by Grunke and Pisoni [1982] and the crucial acoustic information was in F_2 rather than F_3, we expected the auditory salience of the initial transition to be greater than the initial transitions in both Grunke and Pisoni's [1982] /ba/–/da/ continuum and the /ra/–/la/ continuum used by Best et al. [1989]. The experiment was thus designed to shed further light on the extent of the speech listener's advantage in perceiving initial formant transitions.

Experiment 1 also had a methodological goal: to learn whether the experimental procedure introduced by Bailey et al. [1977] produces results comparable to the more complicated procedures employed by later researchers. Rather than attempting to induce a 'perceptual set' in the listeners, we relied on the fact that some listeners spontaneously hear sine-wave replicas as speech and others do not.

Method

Subjects

Seventeen undergraduate students (7 female, 10 male) at Indiana University, Bloomington, participated in the experiment for partial course credit in introductory psychology. None of the listeners reported a history of speech or hearing disorders at the time of testing.

Materials

We synthesized an eleven-step continuum of sine-wave replicas of speech using natural productions of /wa/ and /ya/ as models for the continuum endpoints (the synthesis parameters are shown in table 1). The onset frequency of the sine-wave analog of F_2 varied from 750 to 1,850 Hz. The changes in F_2 onset frequency were calculated as equal intervals in Bark units and then converted into hertz using the formula published by Schroeder et al. [1979]. After a transition of 75 ms, the frequency of the F_2 analog reached 1,334 Hz and then gradually rose to 1,370 Hz during the 175-ms steady-state vowel. The F_2 transition was the only property of the stimuli which varied across the continuum. The F_1 analog also had a 75-ms transition from 371 to 723 Hz followed by a slow change over 175 ms to 797 Hz. The F_3 analog was steady-state throughout at 2,722 Hz.

Procedure

Listening sessions were conducted on-line using a PDP 11/34 computer at the Speech Research Laboratory at Indiana University. Listeners were run in groups of 6 or fewer.

We used a standard identification paradigm in which listeners were asked to label stimuli presented in isolation. Following Bailey et al. [1977], we did not attempt to induce a perceptual set (speech or nonspeech) as has been done by other researchers, but rather attempted to assess each listener's spontaneous experience of the stimuli. (We will point out evidence later which shows that the speech/nonspeech distinction using this method is not simply a function of the listener's attentiveness or seriousness in the task.) During a training phase, only the endpoint stimuli were presented (with an interstimulus interval of 2,000 ms). Listeners were told only that they would hear two types of sounds and that they were to learn which button they should press after each sound. After listeners had made a response on a trial, a feedback light was turned on above the button that they should have pressed. Training continued until all listeners in a group had reached or exceeded 90% correct on a block of 20 trials, which usually only required about

Downloaded by:
Univ of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

Table 1. Synthesis control parameters used to create the sine-wave replica stimuli used in the experiments

	Time, ms:	Experiment 1			Experiments 2 and 3		
		0	75	250	0	130	250
	F_1 Frequency, Hz	371	723	797	741	741	485
	Amplitude, dB	60	60	60	60	60	45
No. 1	F_2 Frequency, Hz	**750**	1,334	1,370	1,257	1,257	**1,889**
	Amplitude, dB	55	55	55	50	50	50
No. 11	F_2 Frequency, Hz	**1,850**	1,334	1,370	1,257	1,257	**778**
	Amplitude, dB	55	55	55	50	50	50
	F_3 Frequency, Hz	2,722	2,722	2,722	2,565	2,565	2,565
	Amplitude, dB	43	43	43	43	43	43

Stimulus No. 1 in experiment 1 was modeled after a natural production of /wa/ while stimulus No. 11 was modeled after a production of /ya/. Stimulus No. 1 in experiments 2 and 3 was modeled after a natural production of /ay/ while stimulus No. 11 was modeled after a production of /aw/.

20–40 trials total. Next, listeners were asked to write down their subjective impression of the stimuli and their classification criteria. During the subsequent generalization phase of the experiment, all stimuli from the series were presented and the feedback lights continued to mark the correct response for presentations of the endpoint stimuli. The stimuli were ordered so that each stimulus followed every other stimulus an equal number of times. Finally, listeners completed a written questionnaire in which they were asked again for a subjective impression of the stimuli and their basis for classifying the stimuli.

Results

The data were sorted into three categories based on the listeners' descriptions of the stimuli before and after the generalization phase. Listeners who reported speech percepts before and after the generalization test were classified as 'speech' listeners (n = 5). Those who reported nonspeech percepts before and after the generalization test were classified as 'nonspeech' listeners (n = 7). Finally, listeners who reported nonspeech percepts before the test and speech percepts after the test were classified as 'mixed' listeners (n = 5). No listeners shifted from speech to nonspeech percepts during the generalization phase. Because it was not possible to know when during the test the mixed listeners had shifted from a nonspeech mode to a speech mode, their data were excluded from further analyses.

Figure 1 displays the proportion of left button responses as a function of stimulus and percept (speech versus nonspeech). For speech listeners the left button corresponded to /wa/ and for the nonspeech listeners the left button corresponded to sounds that had a rising frequency glide. The data for each listener were fit with a cumulative probability function using the method of least squares (probit analysis). The slopes and category boundaries (50% crossover point) obtained from the fitted functions were then submitted to separate analyses of variance with percept (speech versus nonspeech) treated as a between-listeners variable. The analyses confirmed that nonspeech listeners could not reliably categorize the stimuli while speech listeners could. The slopes of the labeling functions were less steep

Downloaded by:
Univ of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

for nonspeech listeners (–0.02) compared to the speech listeners (–0.24) [$F(1, 10) = 65.05$, $p < 0.001$]. Only four of the seven nonspeech functions had category boundaries that fell within the stimulus range. These data indicate that speech listeners identified the stimuli very consistently while the nonspeech listeners were quite inconsistent. The difference between speech (6.35) and the four nonspeech (6.33) category boundaries that fell within the continuum was not statistically significant [$F(1, 7) = 0.001$, $p = 0.99$] despite the difference in the slopes of the identification functions.

Figure 2 displays the reaction times of labeling responses. Overall, reaction times were slower for nonspeech listeners (785 ms) than for speech listeners (698 ms), but the difference was not significant [$F(1, 11) = 1.64$, $p = 0.23$]. Although the reaction time function for nonspeech listeners was relatively flat, the function for speech listeners exhibited a peak near the location of the category boundary. This difference in the shape of the functions produced a significant interaction between stimulus number and percept [$F(10, 110) = 2.76$, $p < 0.01$]. The peaked reaction time function for speech listeners is similar to the reaction time function reported by Pisoni and Tash [1974] for a VOT continuum.

Figure 3 displays labeling responses as a function of the preceding stimulus rather than as a function of the stimulus actually labeled averaged across responses to stimuli 4 through 8 and across listeners within groups. Recall that the stimuli were presented in a pseudorandom order such that each token appeared in the context of every other token an equal number of times. We calculated context functions for these stimuli because context effects are known to be greatest for stimuli near the boundary between perceptual categories. Regression lines were fit to each listener's context function and the regression coefficients were analyzed in an analysis of variance. Positive slopes indicate contrast effects; negative slopes indicate assimilative effects. There was a slight contrast effect for nonspeech listeners (0.004) and a slight assimilative effect for speech listeners (–0.004). However, the effects were small and the difference between groups was not significant [$F(1, 10) = 1.49$, $p = 0.25$].

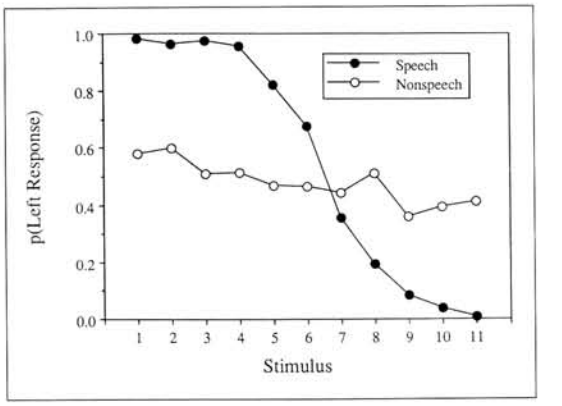

Fig. 1. Labeling functions obtained in experiment 1. Solid dots and connecting lines represent data for listeners reporting speech percepts; open dots and connecting lines represent data for listeners reporting nonspeech percepts.

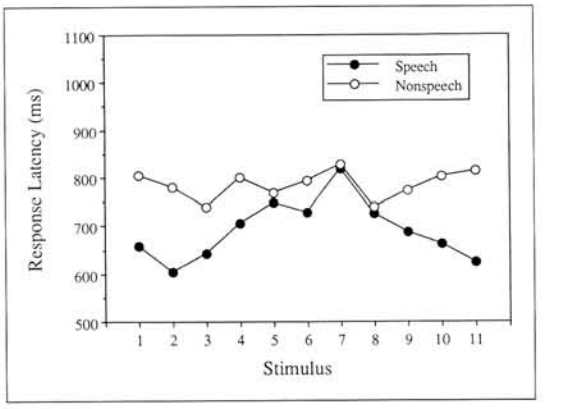

Fig. 2. Reaction times for labeling responses in experiment 1.

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

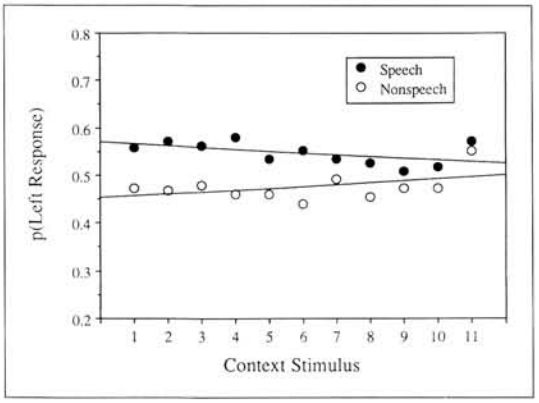

Fig. 3. Context effects for labeling responses in experiment 1. Each point represents the averaged proportion of trials that the left response button was pressed as a function of the preceding stimulus. Each set of points has a corresponding best-fitting line determined by linear regression.

Discussion

The results of experiment 1 are very similar to experiments reported earlier by Grunke and Pisoni [1982] and Schwab [1981] for stop consonants, and by Best et al. [1989] for liquids cued by an F_3 transition. These new results show two things: (1) nonspeech listeners have great difficulty in using initial transitions even when they are quite long and low in frequency, and (2) the Bailey et al. [1977] method of listeners self-selection yields results which are comparable to results obtained using complicated methods to induce speech or nonspeech percepts.

Schwab [1981] noted that nonspeech listeners' inability to categorize complex stimuli with differences only in the initial portion of an otherwise identical pattern suggests the operation of greatly exaggerated versions of low-level auditory masking such as the upward spread of masking (where F_1 masks F_2) and/or backward masking (the steady portion at the end of the simulus masks the transitions). Nygaard and Eimas [1990] and Schwab [1981] also reported effects reminiscent of psychophysical masking.

The automaticity hypothesis predicts that when an acoustic parameter is not psychoacoustically salient speech listeners will outperform nonspeech listeners. However, the lack of a context effect or a reliable difference between the reaction times of speech and nonspeech listeners in experiment 1 suggests that we should either reject the automaticity hypothesis or reassess the predictions of that hypothesis. But, in this first experiment, nonspeech listeners' performance was so random that some interesting and potentially important differences between speech and nonspeech perception may have been obscured.[2] If the automaticity hypothesis is correct, we would expect to find the predicted differences between speech and nonspeech listeners when nonspeech listeners can label the endpoints of the continuum as accurately as speech listeners can. We tested this hypothesis in a second experiment.

[2] If, as these results suggest, nonspeech listeners are unable to attend to certain aspects of the speech signal one wonders how children can acquire such speech sounds in the first place. This issue is beyond the scope of the present article but we can mention some possibilities. It may be that listening in a speech mode engages a special perceptual mechanism and that all a child needs to do to have the speech mode advantage that was demonstrated in this experiment is recognize that a signal is speech (i.e. to engage a special speech mode of perception). On the other hand, the lack of auditory salience illustrated in this experiment may have resulted from the fact that the signals were acoustically impoverished. Studies with animals [Kluender et al., 1987, and others] suggest that naturally produced CV syllables can be categorized without a special perceptual mechanism.

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

Experiment 2

Experiment 2 tested the automaticity hypothesis' predictions that speech perception would be faster and less affected by context than nonspeech perception when the differences between stimuli in a sine-wave replica continuum were salient to nonspeech listeners. Grunke and Pisoni [1982] and Schwab [1981] found that nonspeech listeners could categorize sine-wave analogs of VC syllables even though they could not categorize CV syllables. Therefore, we expected that nonspeech listeners would be better able to categorize stimuli which were roughly the mirror images of the stimuli used in experiment 1, and thus there would be a closer correspondence between speech and nonspeech labeling functions. The automaticity hypothesis predicts that when nonspeech listeners can label the stimuli in a continuum as consistently as speech listeners, their performance will nonetheless differ from speech listeners in two respects: nonspeech listeners will categorize the stimuli more slowly than speech listeners, and nonspeech listeners will show greater context effects than speech listeners.

Method

Subjects

Forty-six undergraduate students (22 female, 24 male) at Indiana University, Bloomington, participated in the experiment for partial course credit in introductory psychology. None of the listeners reported a history of speech or hearing disorders at the time of testing and none participated in the first experiment.

Materials and Procedure

We synthesized an eleven-step continuum of sine-wave replicas of speech with endpoints modeled after natural productions of /ay/ 'eye' and /aw/ 'ou(t)' (table 1). The stimuli were 250 ms in duration and had a steady-state portion (130 ms) and a transition portion (120 ms). The only change across the continuum was in the F_2 transition at the end of the stimuli (rising for /ay/ and falling for /aw/). As with the materials for experiment 1, the F_2 offset frequencies were calculated as equal intervals in Bark units and then converted to hertz. All aspects of the procedure were identical to experiment 1.

Results

Listeners were sorted into three groups according to the same criteria employed in experiment 1: 'speech' (n=15), 'nonspeech' (n=7), and 'mixed' (n=24).[3] As in experiment 1, we based our interpretation of results on the performance of the speech and nonspeech listeners only.

Figure 4 displays the proportion of left button responses as a function of stimulus number and percept. The stimulus which was assigned to the left button was the sine-wave replica of /ay/. Identification data for each listener were fit with probit functions, and the resulting slope and crossover parameters were entered into separate analyses of variance, treating percept as a between-listeners variable. The average speech (5.05) and nonspeech (5.17) crossover points were not significantly different [$F(1, 20) = 0.15$, $p = 0.70$]. Although there was a trend for the slope of the labeling functions to be steeper for speech (–0.31) than nonspeech (–0.23), this difference was not significant [$F(1, 20) = 2.22$, $p = 0.15$]. Therefore, labeling performance was nearly equivalent for the two groups. Both groups were

[3] We do not understand why this continuum was perceived as speech by so many of the listeners. In experiment 1 the ratio of nonspeech listeners to listeners who heard the stimuli as speech at some point during the experiment (the 'speech' and 'mixed' listeners) was 7/10, while in this experiment the ratio was 7/39. Whether this difference in listeners' experience of the stimuli is related to the salience of the F_2 transition is beyond the scope of this article. What is important for our purpose is that there were some genuinely nonspeech listeners.

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

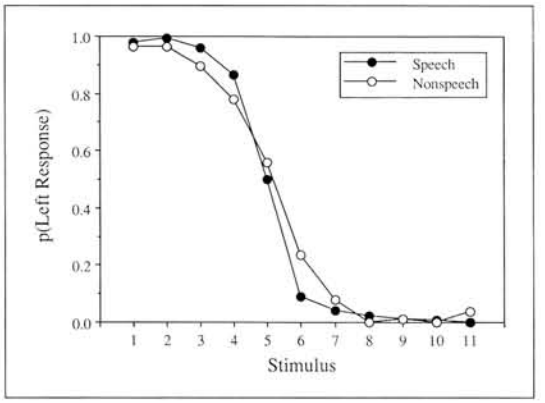

Fig. 4. Labeling functions obtained in experiment 2.

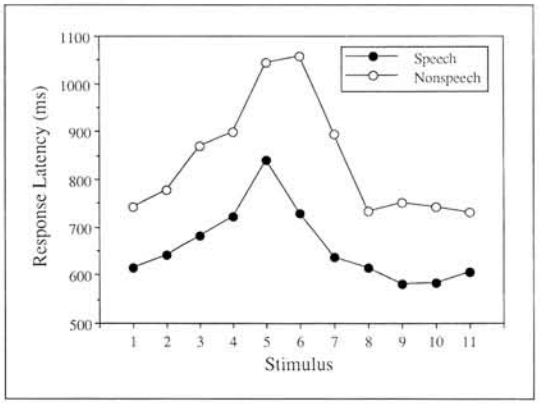

Fig. 5. Reaction times for labeling responses in experiment 2.

able to partition the stimulus series into two relatively discrete perceptual categories with only a small region of ambiguity. One conclusion that this result suggests is that nonspeech listeners were just as careful and as attentive as were the speech listeners. Because we did not attempt to bias listeners toward a speech or nonspeech mode of perception it is important to note that the listeners who (for whatever reason) heard the stimuli as nonspeech categorized the stimulus continuum just as consistently as did the speech listeners.

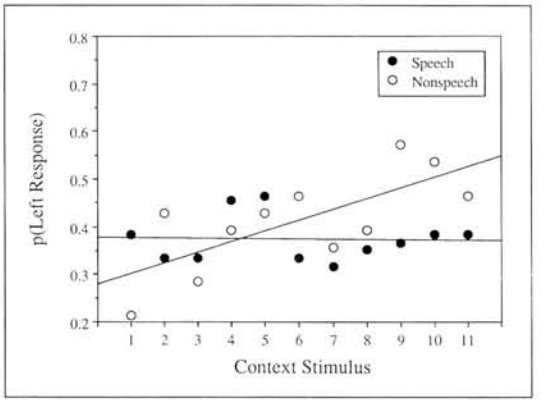

Fig. 6. Context effects for labeling responses in experiment 2. Each point represents the average proportion of trials that the left response button was pressed as a function of the preceding stimulus. Each set of points has a corresponding best-fitting line determined by linear regression.

Figure 5 shows reaction time data for speech and nonspeech listeners. Nonspeech listeners were slower to identify the stimuli than speech listeners [$F(1, 20) = 10.09$, $p < 0.01$]. Given the observed similarities between the two groups' labeling functions, this reaction time difference is not merely a reflection of the nonspeech listeners' inability to categorize the stimuli (as we might have hypothesized for a reaction time difference in experiment 1), but rather appears to reflect a difference in the automaticity of perceptual processing. The token main effect was significant [$F(10, 200) = 22.34$, $p < 0.01$]; responses to tokens near the category boundary were slower than responses to tokens at the endpoints of the continuum. There was also an interaction between percept and token [$F(10, 200) = 2.41$, $p < 0.01$] which appears to reflect the broader peak in the nonspeech listeners' reaction time function.

Figure 6 shows the average influence of the immediately preceding token on identification

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

for the speech and nonspeech listeners. As in experiment 1, context functions were calculated from each listener's responses for the five stimuli in the middle of the continuum (4–8) and simple regression lines fit to the individual data. The slopes from these regression lines were entered into an analysis of variance that treated percept as a between-listeners variable. The average slope for speech listeners (–0.00007) was significantly smaller than the average slope for nonspeech listeners (0.0224) [$F(1, 20) = 8.52$, $p < 0.01$]. Thus, while speech listeners were not influenced by preceding context, nonspeech listeners exhibited a contrast effect.

Discussion

Both of the predictions of the automaticity hypothesis tested in experiment 2 were borne out in these data. The nonspeech listeners identified the stimuli more slowly than the speech listeners. Given the almost random categorization performance of the nonspeech listeners in experiment 1, one could have attributed a reaction time difference in that study to the nonspeech listeners' uncertainty about the judgements they had been asked to make. However, in this experiment, the two groups of listeners had very similar labeling functions. Thus, a stronger case can be made from these results that the reaction time difference reflects a processing difference that is consistent with the automaticity hypothesis. Also, as predicted by the automaticity hypothesis the nonspeech listeners in experiment 2 exhibited a stronger contrast effect than did the speech listeners.

These results show that speech listeners have well-developed, sharply defined categories, but do not necessarily show that the categorization process is automatic. The automaticity hypothesis not only predicts that nonspeech listeners' identification performance will be slower and more sensitive to auditory contrast, but it also predicts that speech listeners will display a functional loss of auditory information as compared to nonspeech listeners. This prediction was tested in a discrimination experiment.

Experiment 3

Method

Subjects

Sixty-six undergraduate students (31 female, 35 male) at Indiana University, Bloomington, participated in the experiment for partial course credit in introductory psychology. None of the listeners reported any history of speech or hearing disorders at the time of testing and none had participated in the previous experiments.

Materials and Procedure

The eleven-step /ay/–/aw/ continuum used in experiment 2 was used in this experiment. Each experimental session was composed of the following parts. First, listeners identified the endpoint tokens in a training phase with visual feedback as in the first two experiments. Following this, listeners wrote their descriptions of the stimuli and their classification criteria. Then, the endpoint stimuli were presented with feedback for identification in a paired comparison, or AX, format. On each trial of this familiarization phase, tokens were presented with an interstimulus interval of 500 ms. After the presentation of the pair of stimuli, listeners were required to identify the first and second tokens using the button labels which they had learned in the training phase. During a subsequent identification test, all possible two-step pairs, as well as pairs in which the stimuli were the same, were presented in random order for identification responses. Next, the same procedure was repeated, but listeners judged whether the stimuli in each pair were the same or different. This discrimination test was also preceded by familiarization trials, again with feedback, using just the endpoint stimuli. In both the identification task and the discrimination task, each of the nine 'different' pairs was presented 8 times for a total of 72 trials, and each of the 11 'same' pairs was presented four times for a total of 44 trials. Finally, listeners filled out a questionnaire in which they again wrote their subjective impressions of the stimuli and their classification criteria.

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

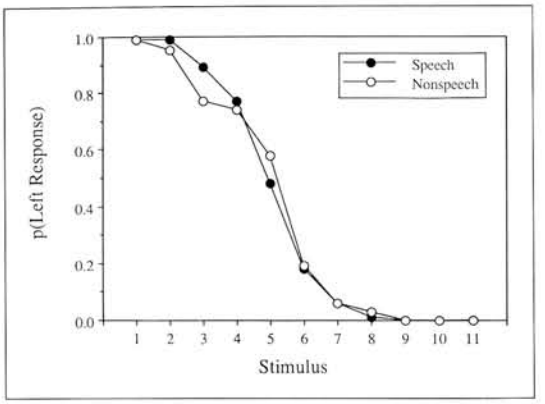

Fig. 7. Labeling functions obtained in experiment 3.

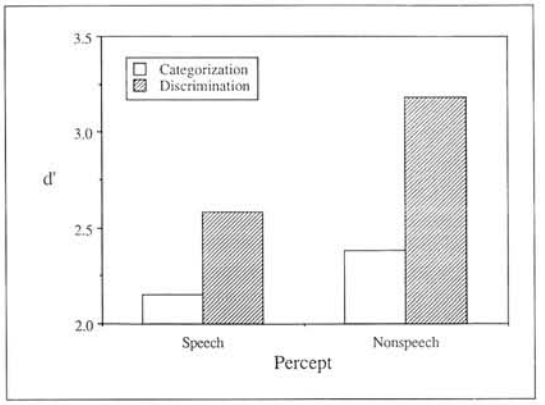

Fig. 8. d' data obtained in experiment 3 averaged across all stimuli. Open bars represent data obtained from identification judgments; striped bars represent data obtained from discrimination judgments.

Results

Listeners were sorted into three groups based on the written descriptions of the stimuli and their classification criteria: 'speech' (n=25), 'nonspeech' (n=17), and 'mixed' (n=24). As in experiments 1 and 2, the data of the mixed group were not included in subsequent analyses. Additionally, because of the somewhat complicated protocol several listeners seemed to have misunderstood one or another of the tasks, therefore 7 listeners (5 nonspeech, 2 speech) were excluded from further data analyses because they performed poorly (<90% correct with the endpoint stimuli) in either the training or familiarization phases in this experiment. Thus, data from 23 speech listeners and 12 nonspeech listeners were analyzed.

Figure 7 shows the identification data for speech and nonspeech listeners. Although the slopes of the functions are less steep than for the corresponding conditions in experiment 2, their overall form is similar to the functions found in that experiment. There was a trend for the crossover point to be higher on the stimulus series for the speech group (5.05) than for the nonspeech group (4.78), but the difference was not significant [$F(1, 34) = 2.15$, $p = 0.15$]. The difference in slope between speech and nonspeech listeners was significant [$F(1, 34) = 8.27$, $p < 0.01$]. The overall decrease in labeling performance in experiment 3 as compared to experiment 2 probably occurred because of the increased memory load of the two-interval labeling task.

Figure 8 shows average predicted and obtained d' values collapsed across stimuli for speech and nonspeech listeners. Predicted scores were derived from the AX identification test. The main result of this experiment was that nonspeech listeners were more sensitive to stimulus differences than were the speech listeners. An analysis of variance was performed on these data, treating percept (speech versus nonspeech) as a between-listeners factor and task (identification versus discrimination) as a within-listeners variable. There was a significant main effect of percept [$F(1, 33) = 4.23$, $p < 0.05$]. The average d' value for speech listeners (2.37) was smaller than for nonspeech listeners (2.78). There

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

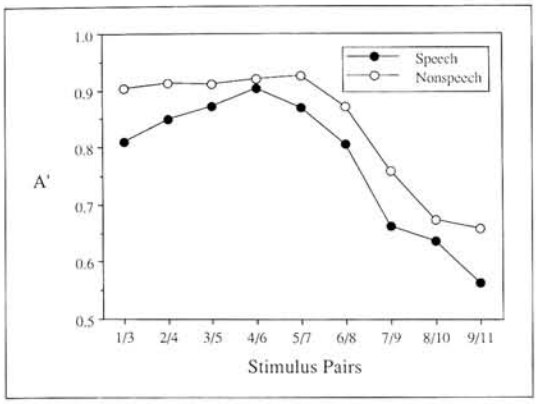

Fig. 9. A' data obtained in experiment 3.

was also a significant main effect for task [$F(1, 33) = 17.41$, $p < 0.01$]. The average d' value in the discrimination task (observed discrimination) was 2.79 while the average d' value in the identification task (predicted discrimination) was 2.23. This is a very common finding in studies of speech perception using a variety of paradigms [Liberman et al., 1961; Healy and Repp, 1982], and it suggests that listeners make use of auditory information during discrimination. The difference between predicted and observed sensitivity was nearly twice as large for nonspeech listeners (0.80) as for speech listeners (0.43), suggesting that speech listeners were making greater use of category labels during the discrimination task [Healy and Repp, 1982]. However, the interaction between percept and task was not significant [$F(1, 33) = 1.6$, $p = 0.21$].

Figure 9 shows discrimination functions for both groups of listeners. A' (a nonparametric analog of d') was used because we wanted a measure of discrimination performance which takes into account response bias and could be calculated using a relatively small number of observations on each stimulus pair. Discrimination functions for each listener were entered into a analysis of variance treating percept as a between-listeners variable and stimulus pair as a within-listeners variable. There was a main effect for stimulus pair [$F(8, 264) = 23.13$, $p < 0.001$]. Pairs close to the category boundary were more discriminable than pairs drawn from within categories. This effect is consistent with the results of previous studies and may be due in part to discontinuities in the auditory code for stimuli across the series [Ralston and Sawusch, 1984]. Also, there was a main effect for percept [$F(1, 33) = 6.25$, $p < 0.02$]. The pair-by-percept interaction was not reliable [$F(8, 264) = 0.87$, $p = 0.54$].

Discussion

Note first that these data are further evidence that the difference between speech and nonspeech listeners in the Bailey et al. [1977] paradigm is not that speech listeners are attentive, careful listeners and nonspeech listeners are somehow more sloppy or less focussed on the task. The nonspeech listeners' discrimination performance was better than the speech listeners'.

More importantly, the data confirm the prediction of the automaticity hypothesis that speech listeners would not be as sensitive to small acoustic differences between stimuli in a discrimination task as are nonspeech listeners. Analyses of overall sensitivity and of discrimination functions found differences between speech and nonspeech listeners. These results suggest that speech listeners attend automatically to a categorical level of representation and thus are less able to focus attention on subcategorical auditory properties of the stimuli.

Our results appear to conflict with previous discrimination studies utilizing sine-wave

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

stimuli that found either no reliable difference between speech and nonspeech listeners [Bailey et al., 1977] or better performance for speech listeners [Best et al., 1981, 1989]. These differences may be due partially to procedural differences such as the memory load of the discrimination task, the size of the interstimulus interval, or the nature of the training regimen, but the most important difference seems to be that the nonspeech listeners in the present experiment could reliably categorize the stimuli. The fact that nonspeech listeners in the studies reported by Best et al. [1981, 1989] showed nearly random categorization performance suggests that the differences between stimuli even at opposite ends of the continuum were not salient to the nonspeech listeners, so it should come as no surprise that they performed poorly in two- or three-step discrimination task. The discrimination data reported here are interesting because the speech and nonspeech listeners were equally able to categorize the stimuli.

General Discussion

The hypothesis tested in the present experiments was that speech perception is automatic. We argued that three aspects of auditory perceptual performance can be used to test this automaticity hypothesis. We noted that speed of performance is commonly associated with automatic processing, but by itself is not conclusive evidence [Shiffrin, 1987]. Experiment 2 showed that nonspeech listeners are slower than speech listeners (in a speeded identification task) even when the two groups' labeling functions do not differ.

The automaticity hypothesis also predicted that, if speech is processed automatically, speech listeners would be able to attend to psychoacoustically obscure acoustic dimensions that nonspeech listeners are less able to use. This pattern of results was found in experiment 1 as well as in several studies reported previously [Grunke and Pisoni, 1982; Schwab, 1981; Best et al., 1989].

We also assumed that if speech perception is automatic, categorical rather than peripheral representations will be placed into the focus of the speech listener's attention. This assumption led us to predict that context would have greater impact on perception in a nonspeech mode than on perception in a speech mode. As predicted, when nonspeech listeners' performance was not random (experiment 2), they showed more sensitivity to context. This finding is not unusual. Siegel and Siegel [1977], for example, observed larger context effects for nonmusicians than musicians. However, this is not the best evidence that speech perception is automatic, because it may merely reflect the fact that speech listeners had well-defined categories to use in labeling the stimuli and nonspeech listeners did not.

Better evidence for the automaticity hypothesis comes from a test of the speech listener's ability to disregard the categorical level and attend to the physical properties of the signal. The results of experiment 3 suggested that listeners in a speech mode of listening cannot attend to subcategorical details of the signal with the same proficiency as listeners in a nonspeech mode. This suggests that recourse to categorical labels in the speech mode is not optional, but that the categorical processing of speech is obligatory. Our argument is that this obligatoriness is a mark of automaticity, of a cognitive process which, once begun, must run its course.

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM

Acknowledgements

Many thanks to Allard Jongman, Richard Pastore, and Joan Sereno for their comments and suggestions on an earlier version of the article. Thanks also to Catherine Best and Arthur Samuel for their thoughtful reviews of the penultimate version of the article. We had fruitful (and enjoyable) discussions concerning this work with Steven Goldinger, John Logan, John Mullennix, David Pisoni, James Sawusch, Richard Shiffrin, and Van Summers. Denise Beike helped us conduct the experiments. This research was supported by NIH Training Grants to Indiana University (DC-00012-11) and UCLA (DC-00029-1) and NIH Grant R29 DC01645-01.

References

Bailey, P. J.; Summerfield, Q.; Dorman, M.: On the identification of sine-wave analogues of certain speech sounds. Haskins Lab. Status Rep. Speech Res., SR 51/52, pp. 1–25 (Haskins Laboratories, New Haven 1977).

Best, C. T.; Morrongiello, B.; Robson, R.: Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception Psychophysics *29:* 191–211 (1981).

Best, C. T.; Studdert-Kennedy, M.; Manuel, S.; Rubin-Spitz, J.: Discovering phonetic coherence in acoustic patterns. Perception Psychophysics *45:* 237–250 (1989).

Brière, E. J.: An investigation of phonological interference. Language *42:* 768–798 (1966).

Flowers, J. H.; Polansky, M. L.; Kerl, S.: Familiarity, redundancy, and the spatial control of visual attention. J. exp. Psychol. hum. Percept. Perform. *7:* 157–166 (1981).

Grunke, M. E.; Pisoni, D. B.: Some experiments on perceptual learning of mirror-image acoustic patterns. Perception Psychophysics *31:* 210–218 (1982).

Healy, A. F.; Repp, B. H.: Context independence and phonetic meditation in categorical perception. J. exp. Psychol. hum. Percept. Perform. *8:* 68–80 (1982).

Kluender, K. R.; Diehl, R. L.; Killeen, P. R.: Japanese quail can learn phonetic categories. Science *237:* 1195–1197 (1987).

Liberman, A. M.; Harris, K. S.; Kinney, J. A.; Lane, H.: The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. exp. Psychol. *61:* 379–388 (1961).

Nygaard, L. C.; Eimas, P. D.: A new version of duplex perception: evidence for phonetic and nonphonetic fusion. J. acoust. Soc. Am. *88:* 75–86 (1990).

Pisoni, D. B.: Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception Psychophysics *13:* 253–260 (1973).

Pisoni, D. B.; Tash, J.: Reaction times to comparisons within and across phonetic categories. Perception Psychophysics *15:* 285–290 (1974).

Ralston, J. V.; Sawusch, J. R.: Perception of sine wave analogs of stop consonant place information. J. acoust. Soc. Am. *76:* suppl. 1, M 7 (1984).

Remez, R. E.; Rubin, P. E.; Pisoni, D. B.; Carrell, T. D.: Speech perception without traditional speech cues. Science *212:* 947–950 (1981).

Schroeder, M. R.; Atal, B. S.; Hall, J. L.: Objective measure of certain speech signal degradations based on masking properties of human auditory perception; in Lindblom, Öhman, Frontiers of speech communication research, pp. 217–229 (Academic Press, New York 1979).

Schwab, E. C.: Auditory and phonetic processing for tone analogs of speech; doct. diss. State University of New York at Buffalo (unpublished, 1981).

Shiffrin, R. M.: Attention; in Atkinson, Hernstein, Lindsey, Luce, Steven's handbook of experimental psychology, vol. 2, pp. 739–811 (Wiley, New York 1987).

Shiffrin, R. M.; Schneider, W.: Controlled and automatic information processing. II. Perceptual learning; automatic attending, and a general theory. Psychol. Rev. *84:* 127–190 (1977).

Siegel, J. A.; Siegel, W.: Absolute identification of notes and intervals by musicians. Perception Psychophysics *21:* 143–152 (1977).

Tomiak, G. R.; Mullennix, J. W.; Sawusch, J. R.: Integral processing of phonemes: evidence for a phonetic mode of perception. J. acoust. Soc. Am. *81:* 755–764 (1987).

Downloaded by:
Univ.of California Berkeley
169.229.32.36 - 1/13/2015 8:29:49 AM