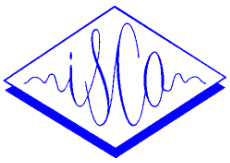

ISCA ITRW on
Spoken Word Access Processes (SWAP)
Nijmegen, The Netherlands
May 29-31, 2000

CONTEXT EFFECTS ON SENSITIVITY AND RESPONSE BIAS

John Kingston
Linguistics Department, University of Massachusetts, Amherst

ABSTRACT

A three-part experiment compared the contrastive effect of preceding liquid consonants on /d-g/ categorization and discrimination. The first and second parts compared the effect of a [l] or [r] with a liquid ambiguous between [l] and [r]. In one, it was much more likely that [l] would follow the preceding vowel than [r]; in the other, the transitional probabilities were reversed. In the third, the context was a 1 ERB-wide energy band centered between [l]'s F3-F4 offset frequencies or [r]'s F2-F3 offset frequencies. Transitional probabilities predict the identity but don't change the ambiguous liquid's perceptual representation, so they may shift the /d:g/ category boundary but not change sensitivity to stimulus differences. The energy band's frequency, however, could also change the consonant's perceptual representation and with it sensitivity. If only the category boundary shifts when listeners actually hear [l] or [r], knowledge of how the liquid affects the pronunciation and acoustics of the following stop has been applied top-down in identifying it. If sensitivity is also affected, the liquid has also changed the perceptual representation constructed during bottom-up processing.

1. INTRODUCTION

Context shifts category boundaries contrastively [14]. Two well-known and related are the shifts in the /d:g/ or /t:k/ category boundary along an F3 onset continuum following [l] or [s] vs [r] or [ ]. Compared to when there is no preceding context, listeners divide this continuum in favor of "g/k" following [l/s] but in favor of "d/t" following [r/] [7-9]. This shift is contrastive because the categories characterized by a low F3 onset frequency, /g,k/, are favored after a sound whose F3 is high, [l], or one in which energy is concentrated at high frequencies, [s], and vice versa.

It has been argued that the boundary shifts compensate for the expected coarticulatory effects of the preceding liquid or fricative on the stop [4,5,7-9]. That is, listeners expect a stop's F3 onset frequency to be higher following [l/s/, and they discount some of its elevation and respond "g/k" more often. In this interpretation, the contrast is between the identity of the context and target segments. This interpretation is plainly an argument for top-down feedback from prior knowledge onto sensory transduction. However, the /d:g/ boundary shifts in the same way when the preceding liquid, [l] or [r], is spoken by a speaker of a different gender or when it is replaced by either a frequency-modulated sine wave that follows F3's trajectory or a constant frequency sine wave that matches the liquid's F3 offset frequency [6]. The listeners in these experiments instead adjusted their criterion for categorizing F3 onset frequency as high or low relative to how high or low energy was concentrated in the preceding sound's spectrum. The listener would not expect the target segment to be coarticulated with the context segment if they're spoken by noticeably different speakers, and can in any case not attribute the acoustic properties heard to the articulations of a single vocal tract. Nor can a single sine wave evoke compensation for coarticulation, no matter how closely it mimics the original liquid's acoustics. These results suggest that the contrast instead arises during bottom-up auditory processing.

However, top-down feedback alone can produce contrastive boundary shifts. They can be obtained with an auditorily ambiguous context, so long as it can be identified from an unambiguous visual signal [4], transitional probabilities [13], or stimulus blocking [3,12].

The experimental conditions manipulate context such that its effects arise either bottom-up during sensory transduction or top-down as prior knowledge feeds back on transduction. Bottom-up context effects are measured with constant frequency 1 ERB-wide energy bands centered between [l]'s F3 and F4 offset frequencies or between [r]'s F2 and F3 offset frequencies [6], and top-down effects by manipulating the transitional probability between an acoustically ambiguous liquid and the preceding vowel [13]. It is not known whether the original liquids' effects arise bottom-up or top-down [4,5,7-9], but comparison of the results obtained when [l] and [r] are present in the signal with those in which the liquids can only be predicted or in which only their distinguishing characteristics are heard should sweep away this uncertainty.

To do so, sensitivity measures were collected in addition to the bias measures reported in previous work. If the sensitivity to an F3 onset frequency difference is greater when the more [d]-like stimulus follows an [r] and the more [g]-like stimulus follows an [l] than in the opposite target-context pairing, then the context segments have altered the bottom-up sensory transduction of the stimuli. On the other hand, if sensitivity is unaffected by the pairing of target and context segments, then the observed boundary shifts are entirely the product of top-down changes in the linguistic weight listeners attach to the products of sensory transduction as a function of context.

2. METHODS

2.1. Stimuli

The stimuli were modeled closely on the productions of two adult male speakers of American English, from

California and Indiana, respectively. They produced a number of tokens of the words "date, gate, tape, cape" in the frame "Say ___ again." The first four formant frequencies and their bandwidths were extracted via LPC analysis (14 coefficients) and averaged across tokens of the same type for use as models for synthesis.

A 10-step [d-g] continuum was constructed by varying F2's onset frequency in equal steps from 1500-2000 Hz and F3's onset frequency in equal steps from 2700-2100 Hz. The two formants changed over the ensuing 100 ms to their initial target values of 1800 and 2400 Hz. A 20 ms noise burst began at the onset of this transition; at the /d/ end of the continuum, the energy in this burst extended from F2-F4, whereas at the /g/ end it was concentrated only at F2 and F3. These manipulations were all intended to produce high quality tokens of the intended syllables. Other synthesis parameters were the same in all stimuli and were set so as to produce a "date-gate" continuum. The initial consonant in these syllables is the target segment.

The speakers also produced nonsense syllables of the shape [hVL], where the L was either [l] or [r] and the V was one of [a,ɔɪ,ɪ]. The vowels in these syllables differ in their transitional probabilities to [l] vs [r], as calculated from English lemma frequencies [2]. Only 24 lemmas with a token frequency of 583 in 17.9 million words contain a syllable ending in [al] vs 1162 lemmas with a token frequency of 85118 that contain a syllable ending in [ar]. On the other hand, 94 lemmas with a token frequency of 6157 contain a syllable ending in [ɔɪl] vs just 4 with a token frequency of 1271 that contain a syllable ending in [ɔɪr]. Transitional probabilities from [ɪ] to both [l] and [r] are very high: 400 lemmas with a token frequency of 134096 have a syllable ending in [ɪl] and 448 lemmas with a token frequency of 105698 have a syllable ending in [ɪr].

Three sets of preceding [CVL] syllables were synthesized using the average formant frequencies and bandwidths of the [hVL] syllables as models for the [VL] portion. Their [CV] portions differed: [na], [pɔɪ], vs [zɪ], with the initial consonant chosen so that no word was formed when either [l] or [r] was added to the [CV]. The [L] was [l], [r], or a liquid ambiguous between the two. F2 and F3 offset frequencies were 1200 and 2650 Hz for [l] and 1200 and 1600 Hz for [r] after [na]; they were 800 and 2550 Hz for [l] and 1500 and 1750 Hz for [r] after [pɔɪ]; and they were 1000 and 2650 Hz for [l] and 1600 and 1800 Hz for [r] after [zɪ]. F2 and F3 frequencies (and bandwidths) also differed at earlier points between syllables ending in [l] vs [r] in ways intended to produce high quality tokens of the intended syllables. The syllables ending in a liquid ambiguous between [l] and [r] were constructed by using F2 and F3 offset frequencies midway between these extremes.

These three sets of syllables were each attached to the members of the "date-gate" continuum, with their offsets separated from the following onset by a 85 ms stop closure which was voiced throughout.

The transitional probabilities predict that a listener will be more likely to identify an ambiguous liquid as [l] rather than [r] after [ɔɪ] but as [r] rather than [l] after [a]. Because both [l] and [r] are equally likely after [ɪ], the [zɪl/r] syllables serve as a control. These manipulations permit us to measure the top-down effects of the predicted liquid identity on the location of the /d:g/ category boundary [13].

A final set of preceding "syllables" were constructed by passing the [zɪl] and [zɪr] syllables through 1-ERB wide bandpass filters centered on either the midpoint between [l]'s F3 and F4 offset frequencies or the midpoint between [r]'s F2 and F3 offset frequencies. For [l], the midpoint was 3025 Hz and the filter passed energy between were 2855-3205 Hz; for [r], the midpoint was 1700 and the filter passed energy between 1599-1807 Hz. The [l]-centered band of noise was then scaled up to be equal in RMS energy to the [r]-centered band. Finally, these filtered syllables were attached to all the members the "date-gate" continuum.

These energy bands do not sound like speech and thus permit us to measure the bottom-up effects of concentrating energy at high or low frequencies on the location of the /d:g/ category boundary [6] and on sensitivity to stimulus differences. The stimuli constructed with the unfiltered [zɪl] and [zɪr] syllables serve as the controls in this condition as well.

2.2. Procedure

Listeners were adult native speakers of American English recruited from the University of Massachusetts, Amherst undergraduate student body. None reported any hearing or speaking pathology. All were paid for their participation. They identified and discriminated the stimuli.

Stimuli were presented for identification in AXB format, where A and B were the endpoints of the [d-g] continuum and X was any stimulus from the continuum. The listeners' task was to say which endpoint category X belonged to. With respect to their position along the continuum, stimuli were presented as X in a proportion of 1:2:2:3:3:3:3:2:2:1 in each block of identification trials. Each block contained both AXB and BXA orders. In the first and second conditions, these blocks consisted of 132 trials; listeners heard 48 unscored training trails at the beginning of the first of these blocks. In the third condition, a block consisted of 88 trials, (32).

Stimuli were presented for discrimination in AXB format, where A and B were stimuli differing by two steps along the [d-g] continuum in the range between stimuli 3-8. X was always identical to A or B and the listeners' task was to say which. For each AB pair, there are four orders of presentation (AAB, ABB, BAA, BBA), and each order was presented once in a block of discrimination trials. In the first and second conditions, blocks consisted of 144 trials, (48 training). In the third condition, blocks consisted of 64 trials (32 training).

In both identification and discrimination, the ISI was 450 ms. Following the third stimulus, the listeners had 2 s to respond by pressing a labeled button before the next trial was presented. Following the listener's response in discrimination, a light came on for 500 ms above the button corresponding to the correct answer.

Four identification and discrimination blocks were presented in each condition yielding a total of 24 identification responses/listener for each of the four middle stimuli in the continuum and 16 discrimination responses/listener for each of the four two-step intervals between stimuli 3-8. Three two-hour sessions run on different days were required to collect all the responses.

In identification, all three stimuli had the same context segment within a trial, [l], [r], or ambiguous [lr] in the first two conditions and [l] or [r] and [l]-centered band or [r]-centered band in the third, but the context segment varied from trial to trial within a block. Listeners identified both the context and target segments on each trial. In discrimination, the context segment was either the same in all three stimuli and the stimuli differed only in the target segment or the context and target segments both differed. In trials where both segments differed, the more [d]-like target was paired with the [r] context and the more [g]-like target with the [l] context and vice versa.

In both identification and discrimination trials, listeners were instructed to respond as quickly as possible, so that response times could be used as a dependent measure as well identification category or discrimination accuracy. Response times are particularly useful in discrimination of stimuli that differ in both the target and context segment as they estimate sensitivity to differences between stimuli that may be perfectly discriminated [1].

Listeners participated in one of three conditions: (1) [ɔɪ] vs [ɪ], (2) [a] vs [ɪ], and (3) filtered [ɪ] vs unfiltered [ɪ]. In the first two conditions, the liquid following the vowel in the first syllable was either [l], [r], or ambiguous [lr], so these conditions compare the boundary shifts produced by predicted vs observed contexts. The third condition compares the boundary shifts and sensitivity changes caused by nonspeech vs nonspeech contexts with the same frequency differences.

Results are reported here from eight listeners each in the [ɔɪ] vs [ɪ] and filtered [ɪ] vs unfiltered [ɪ] conditions, and from six listeners in the [a] vs [ɪ] condition.

2.3. Predictions

First, more "d" responses are expected following [r] than [l]. More may also be observed after the ambiguous liquid [lr], too, but that trend should be modulated by the quality of the preceding nucleus. Compared to the control nucleus [ɪ], more "d" responses are predicted for the test nucleus [a] and fewer "d" responses for the test nucleus [ɔɪ], because the [ar]'s transitional probability is greater than [al]'s but [ɔɪl]'s transitional probability is greater than [ɔɪr]'s. Third, the filtered first syllables are predicted to produce the same contrast as the unfiltered ones, because they also differ in whether energy is concentrated at high vs low frequencies. Finally, if the context changes the stops' perceptual representations, then listeners should discriminate stimuli in which a more [g]-like stop is preceded by a more [l]-like liquid and a more [d]-stop is preceded by a more [r]-like liquid better or perhaps faster than stimuli in which the stop and liquid are combined in the opposite ways.

3. RESULTS

3.1. Identification

The identification results will be reported by analyzing the total proportions of "d" and "l" responses pooled across the [d-g] continuum.

In the two conditions that contrasted different test nuclei in the first syllable with the control nucleus, [ɪ], the pooled "d" and "l" response proportions obtained from each listener were analyzed in separate repeated measures ANOVAs. The contrasting test nuclei, [a] vs [ɔɪ], were a between-subjects independent variable, and the final liquids, [l] vs [r] vs ambiguous [lr], and test vs control nuclei, [a] or [ɔɪ] vs [ɪ], were within-subjects independent variables.

In response to stimuli in which the control nucleus [ɪ] occurred in the first syllable, listeners responded "d" more often when that syllable ended with the ambiguous liquid [lr] than with either [l] or [r] ([lr], with test nucleus [a]: mean = .550, se = .029; [lr], [ɔɪ]: .551, .025; [l], [a]: .535, .037; [l], [ɔɪ]: .491, .032; [r], [a]: .489, .027; [r], [ɔɪ]: .511, .024). The proportions of "d" responses differed markedly between stimuli with the two test nuclei as a function of the final liquid in the first syllable. When the test nucleus was [a], "d" proportions were very similar across all three final liquids ([l] .528, .022; [lr] .523, .024; [r] .543, .025), but when it was [ɔɪ], "d" responses increased progressively from [l] (.556, .019) to ambiguous [lr] (.596, .020) to [r] (.643, .022). These effects are reflected in significant interactions between Liquid x Test vs Control [$F(2,24) = 9.504$, $p = .001$, MSE = .0130], and Liquid x Test Nucleus [$F(2,24) = 4.826$, $p = .017$, MSE = .0086].

These results also show that the first prediction is not confirmed: listeners do not uniformly respond "d" more often following [r] than [l] (or [lr]). That effect is in fact only robustly observed after the test nucleus [ɔɪ].

A planned comparison of the "d" response proportions after the ambiguous [lr] as a function of the preceding nucleus showed no difference following the control nucleus [ɪ] (.550, .029 for the [a] vs [ɪ] condition and .551, .025 for the [ɔɪ] vs [ɪ] condition). Unexpectedly, more "d" responses were obtained after [ɔɪ] (.596, .020) than [a] (.523, .024). Although not significant, this difference is unexpected because the

higher transitional probability between [ɔɪ] and [l] should lead listeners to hear the ambiguous [lr] as "l," which should in turn produce fewer not more "d" responses.

Only the preceding liquid influenced the proportion of "l" responses significantly; planned comparisons showed that "l" responses differed between all three liquids in exactly the expected way, [l] (.981, .007) > [lr] (.707, .090) > [r] (.022, .006).

In the filtered vs unfiltered condition, the proportion of "d" responses was higher when the first syllable was unfiltered (.525, .021) than when it was filtered (.476, .024) and when that syllable ended in an [r] (.523, .020) rather than an [l] (.477, .026). This is the expected effect of the contrast in the preceding liquid. A repeated measure ANOVA with filtering and preceding liquid as within-subjects independent variables showed that both these effects were significant and that they didn't interact with one another [Filtering: $F(1,7) = 20.601$, $p = .003$, MSE = .0192; Liquid: $F(1,7) = 9.904$, $p = .016$, MSE = .0170]. The proportion of "l" or "l"-like responses was affected only by whether the liquid was [l] or [r] or their filtered counter parts, and not by filtering itself [Liquid: $F(1,7) = 8116$, $p < .001$, MSE = 7.229].

3.2. Discrimination

Discrimination was disappointing. In all three conditions, listeners were less accurate and slower when the stimuli differed only in the F2 and F3 onset frequencies in the following syllable than when they also differed in the identity of the liquid at the end of the preceding syllable. And in the pairs in which the stimuli differed in both dimensions, responses were no more accurate or faster when the more [g]-like initial stop was preceded by a more [l]-like final liquid and the more [d]-like initial stop was preceded by a more [r]-like final liquid than vice versa.

4. DISCUSSION

The listeners in this study did not always give more "d" responses following [r] than [l], cf. [4-9]; this prediction was in fact confirmed only following the test nucleus [ɔɪ] and in filtered vs unfiltered condition. The results reported here also do not replicate the earlier finding [13] that listeners use transitional probabilities to predict the identity of ambiguous segments and in turn use those predictions contrastively in judging adjacent segments. Instead, "d" responses increased most from [l]-[lr]-[r] following [ɔɪ] despite the fact that its transitional probability to [l] is much higher than to [r]. Finally, these results did not confirm the prediction that the perceptual representation of the initial stop in the second syllable of these stimuli is changed through contrast with the preceding liquid in ways that alter sensitivity to stimulus differences. Further data will be collected to discover whether these unexpected results arise from stimulus properties or the procedures.

5. ACKNOWLEDGEMENTS

This work was supported by NIDCD, NIH, through grant DC01708. Ling. Dept., U. Massachusetts, Amherst, MA, 01003, US, jkingston@linguist.umass.edu.

6. REFERENCES

[1] Ashby, F. G. & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *J. Math. Psych.*, **38**, 423-466.

[2] Baayen, R., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX lexical database*. Philadelphia: Linguistic Data Consortium.

[3] Bradlow, A. R. & Kingston, J. (1990). Cognitive processing in the perception of the speech, *J. Acoust. Soc. Am.*, **88**, S56. (Abstract).

[4] Fowler, C. A., Best, C. T. & McRoberts, G. W. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants, *Percept. & Psychophys.*, **48**, 559-570.

[5] Fowler, C. A., Brown, J. M. & Mann, V. A. (1999). Compensation for coarticulation in audiovisual speech perception, *Proc. XIVth Intl. Cong. Phon. Sci.* (pp. 639-642). San Francisco.

[6] Lotto, A. J. & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification, *Percept. & Psychophys.*, **60**, 602-619.

[7] Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception, *Percept. & Psychophys.*, **28**, 407-412.

[8] Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception, *Cognition*, **24**, 169-196.

[9] Mann, V. A. & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception, *J. Acoust. Soc. Am.*, **69**, 548-558.

[10] Massaro, D. W. & Cohen, M. M. (**1983**). Phonological context in speech perception, *Percept. & Psychophys.*, **34**, 338-348.

[11] Massaro, D. W. & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception, *Cog. Psych.*, **23**, 558-614.

[12] Ohala, J. J. & Feder, D. (1994). Listeners' normalization of vowel quality is influenced by 'restored' consonantal context, *Phonetica*, **51**, 111-118.

[13] Pitt. M. A. & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *J. Mem. Lg.*, **39**, 347-370.

[14] Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception, *Psych. Bull.*, **92**, 81-110.