INTERSPEECH 2009 BRIGHTON

Learning and generalization of novel contrastive cues

Meghan Sumner [1]

Department of Linguistics, Stanford University, United States
sumner@stanford.edu

Abstract

This paper examines the learning of a novel phonetic contrast. Specifically, we examine how a contrast is learned – do speakers learn a specific property about a particular word, or do they internalize a pattern that can be applied to words of a particular type in subsequent processing? In two experiments, participants were trained to treat stop release as contrastive. Following training, participants took either a minimal pair decision or a cross-modal form priming task, both of which include trained words, untrained words with a trained rime, and novel, untrained words. The results of both experiments suggest that both strategies are used in learning – listeners generalize to words with similar rimes, but are unable to extend this knowledge to novel words.

Index Terms: phonetic variation, acoustic cues, contrast, word learning, form priming, minimal pair decision

10.21437/Interspeech.2009-131

1. Introduction

Listeners are faced with an incredibly variable speech signal. As listeners, we must filter through this variation to find meaning. Sometimes, this variation is arbitrary (as indexical variation such as gender, age, etc.). Other times, this variation is systematic (e.g., tapping in American English). Various studies have shown that adults and children are sensitive to both types of variation [1]-[4], and this aids accommodation [5],[6]. Nevertheless, two outstanding issues are: (1) How is this information used by the system once learned by the listener, and (2) What is learned – a process, or an fact about a particular word? These questions are critical to our understanding of how variable speech is accommodated. In this paper, we taught native English speakers a novel phonetic contrast, and asked not whether they can learn this contrast, but how this contrast is learned – is this a word-specific property or a generalization made across similar words?

2. Experiment 1

This experiment was designed to address the two issues raised above: Do listeners use a learned phonetic contrast in subsequent processing, and if so, is this contrast generalized to novel, untrained words? Participants were trained to associate unreleased stops with a word-final voiced stop (*e.g.,* [bitˀ] = *bead)*. Once trained, they were presented with a main task that included trained items (e.g., [bitˀ]), untrained items that had a trained rime (e.g., [sitˀ]), and novel items (e.g., [kɑtˀ]). If this learned contrast is used in subsequent processing, participants should map the unreleased versions of trained words to words ending in voiced stops rather than voiceless stops. There are two potential outcomes with respect to generalization. On one hand, if the trained words are treated as being somewhat idiosyncratic in nature, then we would not expect participants to generalize this mapping process for untrained words, related (trained rime) or unrelated (novel word). If, on the other hand, participants internalize a surface pattern and apply it, we would expect to see a consistent mapping of unreleased stops to voiced stops, independent of overlap with training.

2.1. Methods

2.1.1. Participants

Seventeen Stanford University students participated in this study for pay or course credit. All were monolingual speakers of American English. None reported any hearing deficiencies.

2.1.2. Stimuli

A native speaker of Korean with an obvious English accent was recording reading an English word list. The speaker was 23 years old and had been in the United States for two months. The word list contained 60 critical word pairs ending in final stops. The pairs differed in the voicing of the final stop (e.g., *beat – bead*). Forty of these word pairs contained front vowels (*e.g., beat – bead; rip – rib; fate – fade; wet – wed; lack – lag*), and twenty contained back vowels (e.g., *moot – mood; rope – robe; not – nod*). In addition to the critical pairs, 580 additional filler words (some used for unrelated experiments) were included in the word lists. The word lists were automatically randomized, and then checked manually to make sure that no two critical items were adjacent. The words were presented in this way to discourage contrastive stress. The speaker was asked to read each word one at a time at a normal speaking rate and to pause briefly after each word.

Once recorded, words were cut into individual files using Praat [7]. We created a novel contrast by manipulating stop release. Words ending in voiceless stops were used as the base form of the manipulation. As base words, we checked to ensure that at least one repetition of each critical word (*e.g., beat*) contained a released final stop. We then took each base word, saved each as the voiceless counterpart of the pair, and created an intended voiced counterpart by splicing the release burst, and saving that form as the voiced member of the pair (e.g., for the word *beat,* we created [bit] which corresponded to "beat" in training, and [bitˀ] which corresponded to "bead" in training). This manipulation was completed for all 60 words ending in voiceless stops. Korean speech was used for three reasons: (1) participants had little experience with Korean-accented English; (2) in a preliminary production test of 6 Korean speakers, no V/C ratio difference was found before voiced and voiceless stops; and (3) in that same test, Korean speakers of English used different strategies for final stops (e.g., deletion, insertion, and no release), but the unreleased strategy was the most frequent.[1]

2.1.3. Procedure

The experiment consisted of a pre-test, training, post-test, and main task. All training phases used a minimal pair decision task (MPDT). Participants saw two words on a monitor (e.g., *beat – bead*), followed 500 msec later with the auditory presentation of the *unreleased* version of the word pair (e.g., [bitˀ]). They were instructed to press the button corresponding to the word that was produced.

[1]All speakers had obvious accents, but were Stanford graduate students, and spoke English fluently.

Copyright © 2009 ISCA

In the pre-test, participants received 20 unreleased words, along with 60 filler trials containing 20 minimal pairs that either varied in the final sound (e.g., *mine – mime*) and 40 trials containing minimal pairs that varied in the initial sound (e.g., *room – loom*). The 80 trials were repeated in a second block. Both the order of presentation and the order of the words on the monitor were randomized. The training phase consisted of the same 20 pairs from the pre-test and post-test phase. All twenty word pairs (released and unreleased) contained front vowels (e.g., *beat – bead, rip – rib, lack – lag*). Also included were untrained, but similar words (e.g., shared rime as in *wet* – wed) and untrained words (e.g., words with back vowels as in *coat – code*). Each participant received ten blocks of training items. Each block contained both a released variant corresponding to voiceless stops, and unreleased variant corresponding to voiced stops. Feedback in the form of correct/incorrect responses was given on a trial by trial basis. Following the training session, which lasted approximately 40 minutes, participants received the post-test. The post-test was identical to the pre-test.

The results of the training phase are presented in Table 1. The % Correct refers to the percent of responses in which participants chose the voiced final word of the minimal pair when presented with a word ending in an unreleased stop.

	% Correct	
	Unreleased stops (e.g., [bit˺])	**Other final (e.g., mine - mime)**
Pre-test	17.2%	82.6%
Post-test	66.9%	86.3%

Table 1. Percent Correct for the Pre-Test and Post-Test

The initial bias of the participants when hearing words ending in unreleased stops is largely toward that of the word ending in a voiceless stop. The voiceless member of the pair was chosen 82.8% of the time. This number is similar to the percent correct for non-stop final words. The fact that these numbers are not closer to 100% may reflect a general difficulty with the perception of accented speech. The most critical point is that following training, participants learn to associate unreleased stops with the voiced member of the pair. The percent correct here represents both a learned ability to make use of a novel contrast, ($p < .001$)) and residue of the participants' native bias towards perceiving other primary cues (e.g., duration [8]) to signal the voicing of the upcoming stop as shown in the difference in performance in the post-test between Unreleased Stops and Other Final minimal pairs ($p < .001$).

Following training, all participants completed the main experimental task. In this experiment, the main task was a MPDT without feedback. In this task, participants were presented with unreleased versions of all critical items. Three experimental conditions were examined: Trained Word, Trained Rime, and Novel Word. The Trained Word condition included 20 items that were repeated from training (e.g., beat – bead). The Trained Rime condition included 20 items that were not presented in training, but whose rimes were presented (e.g., seat – seed). The Novel Word condition included 20 items that were not presented in training, and whose rimes were also untrained (e.g., *rot – rod*). To ensure that no overlap between novel words and trained words existed, only items containing back vowels were used in the novel condition. In addition to the 60 critical items were an additional 180 minimal pairs, of which 60 differed in the final sound (e.g., *mine – mime*) and 120 differed in the initial sound (e.g., *lane – rain*). None of the fillers contained final stops. Twenty items from each of these conditions were used as baselines for initial- and final-deviating forms for analyses.

If the exhibited learning of a novel contrast is available for use in subsequent processing, we expect to see post-test-like performance for the the Trained Word condition. We might also expect to see high accuracy rates, along with reaction times that match those for final-deviating pairs like *seem – seen.* Whether this effect extends to the Novel Word condition is a question of generalization. The second hypothesis we examine is whether a fact about particular words is learned, or whether a pattern is internalized. By examining a novel contrast that is a general voicing contrast, and not a segment-specific contrast (voiced – voiceless vs. t – d), we can ask whether this process is in fact internalized and applied to other forms that fit the appropriate pattern. This hypothesis would be supported if we see a high percent correct for the Trained Rime and Novel Word conditions.

2.2. Results and Discussion

All of the data were analyzed both for accuracy rates and reaction times. The data from one participant was excluded due to high error rates on control items (42.1%). Reaction times 3 standard deviations above and below the mean for each condition were excluded from all analyses. This large window was included to accommodate potential individual differences in responding to accented speech. The accuracy rates and reaction times are provided in Table 2. Standard deviations are provided in parentheses.

Condition	Percent Correct	Reaction Time
Initial Control (e.g., *fun – sun*)	82.49 (9.21)	827.52 (130.77)
Final Control (e.g., *seen – seem*)	74.12 (7.85)	1010.11 (131.83)
Trained Word (e.g., *beat – bead*)	61.79 (8.12)	1005.93 (132.22)
Trained Rim (e.g., *seat – seed*)	62.09 (9.27)	1089.38 (118.17)
Novel Word (e.g., *rot – rod*)	22.53 (5.21)	998.75 (120.98)

Table 2. Mean accuracy rates and reaction times for Experiment 1.

Accuracy Rates. A single-factor analysis of variance (ANOVA) was conducted. There was a main effect of condition ($F(4,70) = 32.40$, $p < .01$). Performance in the Final Control was better than the three experimental conditions (Trained Word: $F1(1,28) = 4.4286$, $p < .05$; Trained Rime: $F(1,28) = 6.044$, $p < .05$; Novel: $F(1,28) = 123.36$, $p < .001$). Critically, there is no difference in performance between the Trained Word and Trained Rime conditions ($F < 1$), but one does exist between the Trained Word and Novel conditions ($F(1,28) = 30.399$, $p < .001$) and between the Trained Rime and Novel conditions ($F(1,28) = 49.726$, $p < .01$).

Reaction Times. A one-way analysis of variance (ANOVA) was conducted. Consistent with the accuracy rates, the ANOVA showed a main effect of condition ($F1(4,70) = 11.723$, $p < .001$). In general, the time to respond to words that vary initially is much shorter than the time to respond to words that vary finally ($F1(1,28) = 13.826$, $p < .01$). Additional analyses compare the experimental condition to the Final Control condition to examine how forms with a new contrast are processed in relation to other word pairs that are similar in nature. Perhaps the most noticeable result is that the reaction times for the Final Control and the Trained Word and Novel Word conditions are nearly identical (Final Control –

Trained Word: F1(1,28) < 1, p = 0.7873; Final Control – Novel Word: F1(1,28) < 1, p = 0.5469; Trained Word – Novel Word: F1(1,28) < 1, p = 0.7782. This similarity does not hold, though, for the Trained Rime condition, which is slower overall when compared to the other three conditions (Final Control: F1(1,28) = 9.4677, p < .01; Trained Word: F1(1,28) = 7.1682, p < .05; Novel Word: F(1, 28) = 6.3482, p < .05).

Taken together, these results provide us with a better understanding of both subsequent processing and generalization. First, and foremost, there is no evidence of generalization to novel words. The low accuracy rates, or the inability for participants to move from mapping unreleased voiceless stops to voiced stops for novel words, coupled with the stable reaction times for the novel words when compared to the control, suggest that items in the Novel Word condition have not benefited from an internalized process that can be applied to all forms of a particular type.

While we can conclusively say that in this experiment, there is no hint of generalization to novel words, this does not imply that listeners do not generalize the pattern at all. The Trained Rime condition is interesting for a number of reasons. Most importantly, participants are clearly able to extend something they learned about one set of words in training to an untrained set of words in a subsequent task. This result suggests that what is being learned, while not the complete generalization of a process, is not something specific about particular words. In the Same Rime condition, participants not only improve in their ability to map unreleased stops to the voiced member of a minimal pair – they also take more time to do it. Whether this is the result of competition or the slowing of responses due to some other mechanism (in pseudoword repetition this occurs because an item that is clearly not a lexical item seems familiar [9]) is unclear at this point. Regardless, the novel contrast is used in subsequent processing and is also generalized to some degree at a sub-lexical level.

One may argue that the contrast was generalized because participants were already familiar with the task, and the lengthy reaction times reflect a response strategy. The consistency between training and the main task may promote generalization of the contrast. Exp. 2 addresses this concern by examining the subsequent processing of the learned contrast in a paradigm different from the training task – cross-modal form priming.

3. Experiment 2

Form priming has been used to highlight similarities and differences in the phonological makeup of words. Primes and targets that are either identical (e.g., bead – bead) or related (e.g., beat – bead) are typically compared to pairs that are unrelated (e.g., fun – bead). Generally, targets preceded by identical primes are identified faster than targets preceded by unrelated primes. The same is true for targets related to their primes. Critically, while priming exists in both cases, the magnitude is different for identical versus related targets [9]. With this in mind, we can predict that if the novel contrast from training is generalized and unreleased final stops are systematically mapped to voiced final stops, we would expect to find faster response times to a target (e.g., *bead*) when it is preceded by a word ending in an unreleased stop (e.g., [bit˺]) than for the same target preceded by a word ending in a released stop (e.g., [bit]).

3.1. Methods

3.1.1. Participants

Twenty-seven Stanford University students participated in this study for pay or course credit. The remaining sixteen participants were all monolingual speakers of American English. None reported any hearing deficiencies.

3.1.2. Stimuli.

All of the stimuli in Experiment 1 were used. In addition to those stimuli, four additional items per experimental condition were added to this experiment. These items were recorded along with those words for Experiment 1, and were manipulated in the same manner.

3.1.3. Procedure

The procedure for Experiment 2 was identical to Experiment 1 for the pre-test, training, and post-test phases. The difference was in the main task following the training. The results of training for Experiment 2 are presented in Table 3.

	% Correct	
	Unreleased stops (e.g., [bit˺])	**Other final (e.g., mine - mime)**
Pre-test	15.9%	87.2%
Post-test	65.6%	88.9%

Table 3. Percent Correct for the Pre-Test and Post-Test

As in Experiment 1, participants are able to make stop release contrastive after training ($p < .001$), but are not able to completely usurp native biases ($p < .001$). Also as in Experiment 1, all participants completed the training phase immediately followed by the main experimental task. While Experiment 1 used a MPDT, similar to that used in training to examine both subsequent processing of newly learned contrasts and generalization, Experiment 2 used a form priming paradigm. In this paradigm, participants are presented with an auditorily presented prime followed by a visually-presented target. The target was presented 100 msec after the offset of the prime. Participants were instructed that they will hear a word, followed immediately by a word presented on the monitor, and they must decide whether the word on the monitor is a real word or a pseudoword.

One may argue that an auditory-auditory presentation is preferred; however, this introduces an unwanted confound. Research has suggested that listeners are sensitive to minimal acoustic cues, such as release [10]. If an auditory – auditory paradigm was used, any difference between unreleased and released variants of a word may be due to this sensitivity, and not to training. While this is not completely moot in this design, consistency in results across the two experiments in this paper, and the attempt to control this sensitivity, should be sufficient to address this issue.

Three experimental conditions and one control condition were used in this experiment: Trained Word, Trained Rime, Novel Word, and Control. All targets were matched with three different prime types: Identity (repetition of same word: hear [bit˺], see *bead);* Related (vary by a single feature: hear [bit], see *bead*)*;* and Unrelated (prime is not phonologically related to the target: hear [fΛn], see *bead*).

A within-subject design was used, dictating that for each condition, eight items would be used per prime type. Three counterbalanced lists were created to ensure that all targets were preceded by all primes. No item was repeated within a list. Therefore, including controls, each list contained 96 test trials. A total of 192 prime-target filler trials containing real-word targets were added. Half were pseudoword – real word prime – target pairs, and half were real word – real word prime target pairs. This controlled for pseudoword – pseudoword

identity conditions, in order to avoid having the only identity trials be those including real-word primes. For the real – real filler pairs, one-third of the fillers had unrelated primes and targets, one-third had related (varying by a single feature) primes and targets, and one-third had identical primes and targets. In addition to the real-word filler trials, 192 real-word prime – pseudoword target trials were added. Since the task is to make lexical decisions to target, half must be real and half pseudowords. Of these, half contained real-word primes and half contained pseudoword primes. For the real word – pseudoword filler trials, half contained targets that were unrelated to the prime (e.g., fun – kersh) and half contained targets that were related to the prime by varying a feature (e.g., fun – fum). For the 96 trials that were pseudoword – pseudoword trials, one-third were unrelated, one-third were related by a single feature, and one-third were identical. The number and composition of the fillers should reduce the development of strategic responses. The design resulted in a total of 480 trials. The experiment lasted 30 minutes.

3.2. Results and Discussion

Mean reaction times were analyzed for Experiment 2. As is typical in this paradigm, no effects were found with error rates, as participants are typically good at identifying words and pseudowords. Regardless of which word unreleased stops are mapped to, both options are real words. Error rates are included in parentheses, and standard deviations are presented in brackets. The data from three participants was excluded based on high error rates for non-critical items (> 25%). Reaction times less than 500 msec and greater than 2500 msec were excluded from all analyses.

Condition		Reaction Time	Priming Effect
Control:	*Identity*	601.89 (3.3/66.75)	100.75
	Related	639.49 (2.9/61.46)	63.15
	Unrelated	702.64 (2.4/72.86)	–
Trained Word:	*Identity*	576.82 (2.8/73.85)	104.3
	Related	622.00 (3.1/65.11)	59.12
	Unrelated	681.12 (2.4/53.93)	–
Trained Rime:	*Identity*	623.24 (2.6/74.51)	89.87
	Related	673.29 (2.5/69.65)	39.82
	Unrelated	713.11 (1.9/72.36)	–
Novel Word:	*Identity*	665.09 (2.8/75.91)	46.37
	Related	659.69 (2.1/73.81)	51.77
	Unrelated	711.46 (3.0/64.01)	–

Table 4. Mean Reaction Times and Priming Effects for Experiment 2. Error rates/standard deviations are provided in parentheses.

An two-factor ANOVA (Condition X Related) was used to analyze the results. There were main effect of Condition ($F(3,253) = 24.599$, $p < .001$) and Relatedness ($F(3,253) = 16.365$, $p < .001$). The effect of relatedness is likely driven by the overall faster response times for non-control items as compared to control items. Targets preceded by identical primes were recognized more quickly than those preceded by unrelated targets ($F(3,161) = 11.66$, $p < .001$) and related targets ($F(3,161) = 8.572$, $p < .001$). Targets preceded by related targets were also recognized more quickly than targets preceded by unrelated primes ($F(3,161) = 8.69$, $p < .001$). The most striking result, though, is that we once again have a dichotomy between trained components (words, rimes) and novel words. Planned comparisons showed that a significant difference was found between the Identity and Related conditions in the Trained Word and Rime conditions (Trained Word: $p < .001$; Trained Rime: $p < .001$), but not for the Novel Word condition, in which responses to identical and related targets were nearly identical ($p = 0.39$). The fact that we find a difference between these conditions for trained conditions suggests that participants consistently map the unreleased version of the words to voiced sounds. What we see in the [bit'] – BEAD case is true identity priming, and the reduction in priming in the [bit] – BEAD case suggests that these two words are similar, but not identical. This is not the case for the Novel Word condition, however, in which both primes behave as similar, but not identical. These results suggest that participants have learned the contrast at a sub-lexical level, but not as an internalized contrast in voicing.

4. Conclusions

This paper examined the use and generalization of a novel contrast by listeners of English. All participants learned to make a new map for a novel cue – one that is not contrastive in English. Specifically, participants learned to associate an unreleased final stop with a voiced final stop for trained words. The critical question examined in this paper was whether listeners learn something specific about a handful of words, or whether they learn to correlate release with voicing across words, trained and untrained. The results of experiments suggest that while participants do not generalize to all untrained words, they do generalize to untrained words with trained sub-lexical components, like rimes.

5. Acknowledgements

This research is supported by NSF Grant 0720054 made to Meghan Sumner. Special thanks to Jonathan Pelsis and Nicole Fernandez for help with this project.

6. References

[1] Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. **Cognitive Psychology**, *47,* 204-238.

[2] Sumner, M., & Samuel, A.G. (2005). Perception and representation of phonologically-regular variation: The case of final /t/. **Journal of Memory and Language**, *52,* 322-338.

[3] McMurray, B., Aslin, R., Tanenhaus, M., Spivey, M., and Subik, D. (2008). Gradient sensitivity to within-category variation in speech: Implications for categorical perception. **Journal of Experimental Psychology: Human Perception and Performance**, *34,* 1609-1631.

[4] Gow, D. W. (2003) Feature parsing: feature cue mapping in spoken word recognition. **Perception & Psychophysics**, *65,* 575-90.

[5] Bradlow, A. R. and Bent, T. (2008) Perceptual adaptation to non-native speech. **Cognition,** *106,* 707-729.

[6] Lively, S., Logan, J., & Pisoni, D. (1993). Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in new perceptual categories. **Journal of the Acoustical Society of America,** *94,* 1242-1255

[7] Boersma, P. & Weenik, D. (2008). Praat. doing phonetics by computer (Version 5.0.43). Retrieved from http://www.praat.org/

[8] Port, R. F. & Dalby, J. (1982). Consonant/vowel ratio is a cue for voicing in English. **Perception & Psychophysics**, *32,* 141-152.

[9] Wagenmakers, E.-J., Zeelenberg, R., Steyvers, M., Shiffrin, R. M., & Raaijmakers, J. G. W. (2004). Nonword repetition in lexical decision: Support for two opposing processes. **Quarterly Journal of Experimental Psychology: Human Experimental Psychology**, *57,* 1191–1210

[10] Sumner, M. & Samuel, A. (2009). The effect of experience on the perception and representation of dialect variants. **Journal of Memory and Language**, *60,* 487-501.

[11] Deelman, T. & Connine, C.M. (2001). Missing information in spoken word recognition: Nonreleased stop consonants. J**ournal of Experimental Psychology: Human Perception and Performance**, *27,* 656-663.