

Learning Phonemes: How Far Can the Input Take Us?

Jessica Maye and LouAnn Gerken
University of Rochester and University of Arizona

Many studies on developmental speech perception (e.g. Werker & Tees 1984, Kuhl et al. 1992) have documented changes in speech perception that occur during an infant's first year of life. These changes are generally understood to reflect the phonemic structure of the native language (Best et al. 1988, Liberman et al. 1957). There is little research, however, on the phonological abstractness of these initial phonetic categories acquired in infancy. One study by Jusczyk and colleagues (1999) found that 9-month-old infants are sensitive to whether or not a set of sounds shares phonological features, indicating that by this young age infants have already developed featural representations. The question we ask in the present research is whether phonetic categories are *initially* represented as bearing abstract, contrastive features, or whether additional information or experience is required before a learner will develop featural representations.

It might be the case that infants represent speech sounds as bearing potentially contrastive features immediately, as soon as they learn a speech sound category. For example, an infant who learns that /d/ and /t/ are contrastive in their language, might immediately assume that in their language *voicing* is contrastive. This assumption might be helpful in language acquisition, because some speech sounds are more frequent than others. If alveolar sounds are produced more frequently than velar sounds in a given language, then an infant learning that language might have good evidence that /d/ and /t/ are contrastive before they have heard enough examples to know whether /g/ and /k/ are contrastive. Immediately representing speech sound categories in terms of features could enable infants to bootstrap less robust contrasts from those contrasts that are better represented.

However, the above assumption might work against an infant, if the language in question has an asymmetrical phoneme inventory. The language might have a voicing contrast for alveolar sounds, while only having voiceless sounds at labial and velar places of articulation. So hypothesizing that voicing is contrastive in such a language would turn out to be misleading. To avoid this garden path, infants might start with a more conservative hypothesis, and not develop abstract, featural representations until they have good evidence that there are multiple analogous contrasts in the language exhibiting the same feature.

1. Transfer of a Trained Contrast

Previous studies of phonemic contrast learning have found transfer of a learned contrast to untrained stimuli, providing evidence that newly learned contrasts are represented as bearing contrastive features. McClaskey et al. (1983) trained adult English speakers to discriminate a three-way voicing distinction. Although English has only a two-way voicing distinction, these researchers trained subjects to categorize a voicing continuum into three categories (corresponding to pre-voiced, voiceless, and aspirated sounds). After learning to make this three-category mapping, subjects generalized the same three categories to a different place of articulation, without any additional training.

Tremblay et al. (1997) replicated the McClaskey study using ERP measurements. After being trained to categorize a voicing continuum into three categories, subjects' discrimination was tested using an oddball paradigm in which stimuli from one category were presented repeatedly, occasionally interrupted by a single token from a different category. When the oddball stimulus was presented, subjects' ERPs showed mismatched negativity (MMN), demonstrating their ability to discriminate the two categories. Subjects were then presented with stimuli from the same voicing continuum at a new place of articulation. Without additional training, subjects showed an MMN response to oddball stimuli, indicating that they discriminated the new voicing contrast at an untrained place of articulation.

Wang et al. (1999) found generalization of this sort in a somewhat more natural language task. They trained adult English speakers to discriminate various tones in Mandarin Chinese, and then found that subjects could correctly categorize those tones when produced in new words and by a new speaker.

This previous research demonstrates the ability of adult learners to develop representations of newly learned contrasts that are general enough to allow transfer of training to untrained stimuli. However, these studies differ from natural language acquisition: in these studies the subjects were informed about the number of phonetic categories they *should* discriminate, whereas children learning their native language are not told how many phonetic categories are in the language. In natural language acquisition, infants must simply listen to the speech around them, and determine for themselves how many categories they hear. Because of this important difference, these studies may not accurately characterize native language phonetic category learning.

The question we ask in the present research is whether subjects will still show generalized, featural representations of newly learned categories by transferring discrimination from trained to untrained stimuli if they are never explicitly informed about the number of phonetic contrasts in the language.

2. Implicit Acquisition of a Phonetic Contrast

Our present research is, in part, a replication of a prior experiment (Maye & Gerken 2000), in which we found that adult subjects could learn the phonemic contrasts of a new language, purely on the basis of exposure to a particular statistical distribution of phonetic tokens. Because this method of training does not involve explicit instruction regarding the number of phonetic categories to be learned, it provides a means to test the generality with which implicitly learned categories are represented.

2.1. Stimuli

In our previous study, the phonetic contrast we used was between two sounds, created for the purposes of the experiment — /d/ and /D/ — which are both similar to the English ‘d’. The phonetic characteristics of these two d-like sounds are shown in the Appendix, Figure 6. The sound /d/ differs from a typical English ‘d’ in that it is slightly prevoiced, while /D/ differs from a typical English ‘d’ in that the formant transitions from formant onset to vowel nucleus are steeper for ‘d’ than for /D/.¹

These d-like stimuli were created by altering naturally produced syllables (using Kay Elemetrics ASL speech synthesis software, as well as Macromedia SoundEdit 16), and were then re-synthesized into a continuum ranging from /d/ to /D/.²

2.2. Training Distributions

We reasoned that if phonetic category learning in natural language acquisition occurs on the basis of mere exposure to speech sounds, phonetic categories might be inferred on the basis of the statistical distribution of phonetic tokens than an infant hears. Although there is a great deal of phonetic variation in naturally produced speech, presumably the tokens of a particular phonetic category cluster together, while relatively fewer tokens are produced midway between two category clusters. To put this in more concrete terms, imagine a Language A (exemplified by the broken line in Figure 1), in which there is a phonemic voicing contrast. Speech sounds will be produced along an entire continuum of voice-onset times (VOTs), but the most frequently produced VOTs will fall into two clusters (a bimodal distribution), corresponding to voiced and voiceless categories. Now imagine a Language B (the solid line in Figure 1), in which voicing is not contrastive. Once again, there will be much phonetic variation, but the most frequently produced VOTs will form a single cluster (a monomodal distribution). A sensitivity to phonetic distributions, then, could enable infants to use this information to infer whether they are learning Language A or Language B.³

Subjects were adult English speakers, divided into two groups. During the training phase of the experiment, one group of subjects heard a monomodal

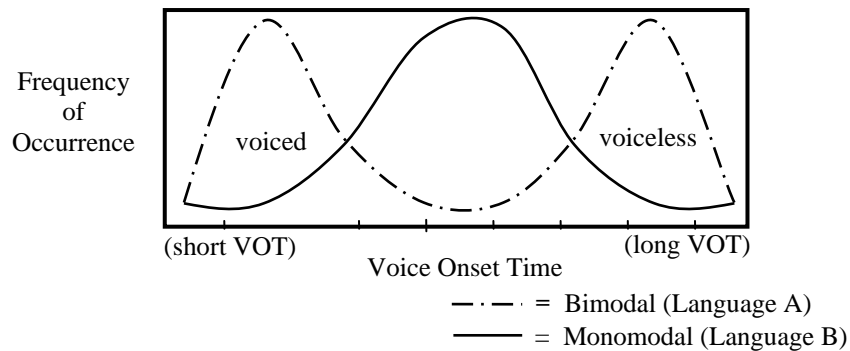


Figure 1. Monomodal and Bimodal Distributions.

distribution of the sounds along the /d~/D/ continuum, while the other group of subjects heard a bimodal distribution of the same stimuli. We predicted that learners would be sensitive to these distributional cues, and use them to infer the phonetic category structure of the language. If this prediction is correct, then during the test phase of the experiment, the bimodal group would be more likely to discriminate between /d/ and /D/ than the monomodal group.

2.3. Results

Our prediction was confirmed, as shown by the test phase discrimination results in Figure 2. The bimodal training group was significantly more likely to discriminate the /d~/D/ contrast than was the monomodal training group.

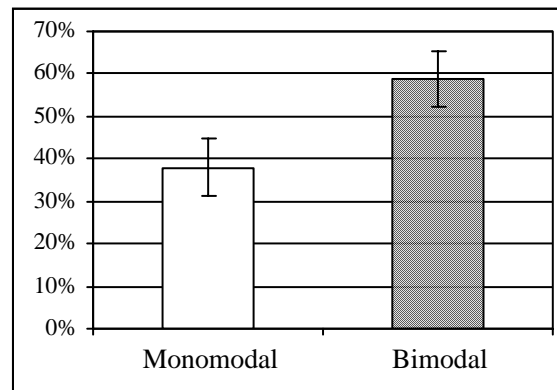


Figure 2. Percent Discrimination of /d~/D/ Pairs (M&G 2000).

These results demonstrate that adults can use the statistical distribution of the sounds produced in a language to infer the language's phonetic category

structure. Because this is a plausible manner in which the phonetic categories of a native language might be learned during infancy, we can use this method of training to test whether newly learned phonetic categories are represented via contrastive features.

3. Experiment

The present study is modeled after Maye & Gerken (2000), but with an added component. Following the test phase in which subjects are tested on their discrimination of the training stimuli, we added a second test phase, testing discrimination of the same type of contrast at a different place of articulation. This experiment is analogous to the previous studies on transfer of a learned contrast (McClaskey et al. 1983, Tremblay et al. 1997, Wang et al. 1999), but here subjects are not explicitly instructed about the number of categories.

3.1. Method

3.1.1. Stimuli

The stimuli were the 18 CV syllables shown in Table 1. There were six consonants, that could each occur in the context of one of three vowels. One consonant continuum (either /d/~D/ or /g/~G/, counterbalanced across subjects) was presented during the training phase (the ‘trained contrast’), along with syllables beginning with the consonants /m/ and /l/, which were included as fillers. The remaining consonant continuum was reserved for the second test phase (the ‘untrained contrast’).

Table 1. “Words” of the Artificial Language.

/da/	/Da/	/ma/	/la/	/ga/	/Ga/
/dæ/	/Dæ/	/mæ/	/læ/	/gæ/	/Gæ/
/dr/	/Dr/	/mr/	/lr/	/gr/	/Gr/

The alveolar stimuli were the same as those used in the previous experiment. The velar stimuli were created to be analogous to the alveolar contrast. They were also re-synthesized from natural speech, and differed from each other along the same phonetic dimensions as the alveolar contrast (see schematic spectrograms, Appendix, Figure 6). The filler stimuli were also re-synthesized, but did not form continua or conform to any particular phonetic distribution.

3.1.2. Participants and Training Distributions

The participants were 64 students at the University of Arizona, who were divided into four groups of 16 each. The groups differed with respect to which type of training distribution they were presented with during the training phase

(monomodal or bimodal), and were counterbalanced with respect to which consonant continuum they heard during training (alveolar or velar).

Figure 3 shows the frequency with which the tokens of the trained-contrast continuum were presented to the groups during training. On the X-axis are the 8 tokens of the continuum, and the Y-axis shows the number of times each token was presented during each block of the training phase. For the monomodal groups, tokens from the center of the continuum were presented most frequently, while for the bimodal groups, tokens near the endpoints were most frequent. The frequency of Tokens 1 and 8 was held constant for both groups of subjects (each was presented only once per block of training) because these two tokens would be used during the test phase, to test subjects' discrimination of the trained contrast.

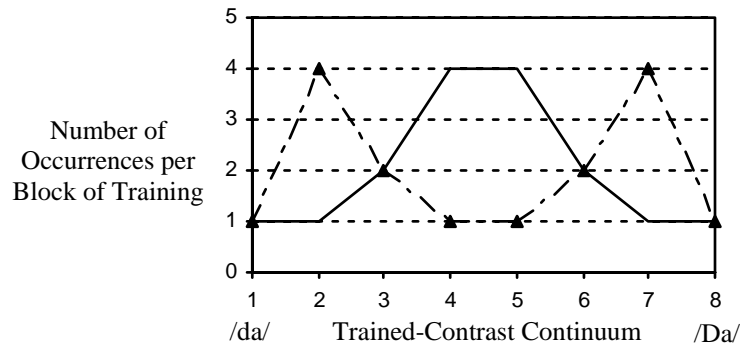


Figure 3. Presentation Frequency of Trained-Contrast Stimuli.

3.1.3. Procedure

Subjects were informed that they would be listening to words from a foreign language, but that they would not know what the words meant. The experiment was conducted in three phases. During the training phase, subjects were asked to simply listen to the 'words' of the language. During this phase, they heard the experimental continuum appropriate for their training group (either alveolar or velar), in each of the three vowel contexts. They also heard four tokens of each of the filler syllables beginning with /m/ and /l/. They were presented with four training blocks, for a total of 9 minutes of training.

The training phase was followed by two test phases. The first test phase assessed discrimination of the trained contrast. Subjects were presented with pairs of syllables and asked to indicate on a response sheet whether each pair was a repetition of a single word in the language, or whether the words in the pair were two different words in the language. Since the two syllables in each pair always shared a vowel, subjects were in essence being asked to make phonemic contrast judgements about the consonants in each pair. Many pairs were filler items, such as /ma/~la/, but the items of particular interest were pairs like /da/~Da/, exhibiting the trained contrast. We predicted that subjects from

the bimodal training groups would be more likely to discriminate these pairs than would subjects from the monomodal groups.

Test Phase 1 was followed by a second test phase, which assessed discrimination of the untrained contrast. As in the first test phase, there were many filler items. During this test, subjects were not presented with any syllables from the trained contrast, but instead heard pairs demonstrating the untrained contrast. Test Phase 2 was the first point in the experiment where subjects heard the untrained contrast. However, if subjects had formed a featural representation of the first contrast, then their discrimination of this second contrast should also reflect their training distributions. That is, the bimodal groups should be more likely to discriminate the untrained contrast than should the monomodal groups.

3.2. Results

3.2.1. Trained Contrast

The results from the first test phase, in which subjects were tested on their discrimination of the trained contrast, are shown in Figure 4. The bars represent the mean percentage of discrimination of the endpoint stimuli in the continuum stimuli that subjects heard during the training phase. For subjects trained on the /d~/D/ stimuli, these results show their discrimination of the /d~/D/ pairs; while for subjects trained on the /g~/G/ stimuli, these results show discrimination of the /g~/G/ pairs.

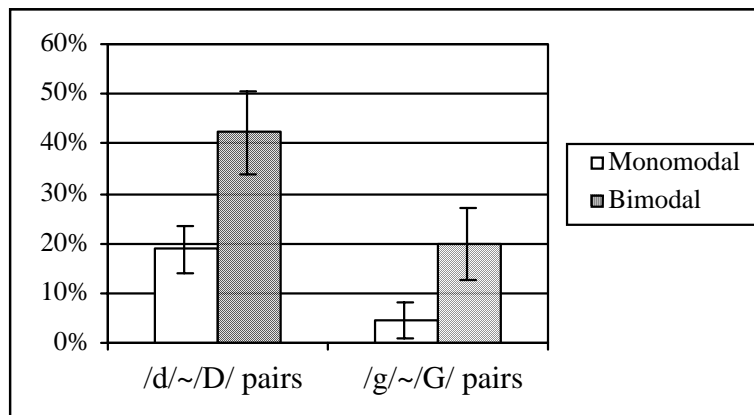


Figure 4. Percent Discrimination of the Trained Contrast, Test Phase 1.

We conducted paired t-tests to evaluate the predicted effects of training distribution on discrimination of the experimental contrasts. For both the alveolar and velar stimuli, subjects trained on a bimodal distribution of the experimental continuum were more likely to discriminate the endpoint tokens

than were subjects trained on a monomodal distribution (alveolar pairs: $t(30) = 2.475$, $p < .01$; velar pairs: $t(30) = 1.87$; $p < .05$).

3.2.2. Untrained Contrast

The results from the second test phase, in which subjects were tested on their discrimination of the untrained contrast are shown in Figure 5. For subjects trained on the /d~/D/ stimuli, these results show their discrimination of the /g~/G/ pairs; while for subjects trained on the /g~/G/ stimuli, these results show discrimination of the /d~/D/ pairs.

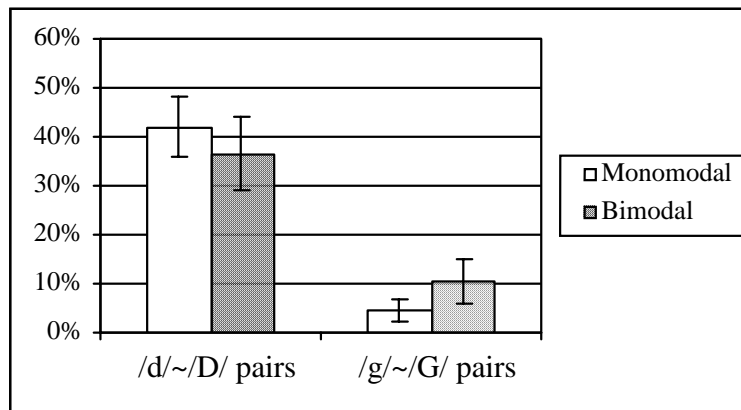


Figure 5. Percent Discrimination of the Untrained Contrast, Test Phase 2.

In contrast with the trained stimuli, subjects trained on a bimodal distribution of the experimental continuum were *not* more likely to discriminate the untrained contrast than were subjects trained on a monomodal distribution (alveolar pairs: $t(30) = .823$, $p = .381$; velar pairs: $t(30) = 1.14$, $p = .265$).

3.3 Discussion

The results from the first test phase replicate our previous finding that mere exposure to a particular phonetic distribution is sufficient to enable learners to form rudimentary phonetic categories. However, the results from the second test phase indicate that the subjects had not formed featural representation of the contrast. In contrast with the previous findings of generalization from trained to untrained stimuli (McClaskey et al. 1983, Tremblay et al. 1997, Wang et al. 1999), in this study where subjects were not instructed as to how many categories there were, training on one set of stimuli did not generalize to untrained stimuli.

There are a number of reasons why subjects may not have developed generalized, feature-based representations in this experiment. One possible

explanation is that subjects in the bimodal groups had not mastered the trained contrast. Evidence from the perceptual learning literature suggests that training on one set of stimuli will not transfer to new analogous stimuli, unless subjects have achieved mastery of the trained contrast (Liu & Weinshall 2000). In our experiment, the bimodal groups only achieved 40% (alveolar pairs) and 20% (velar pairs) discrimination of the trained contrasts, indicating that they had not yet achieved mastery. It is possible that we would find generalization if subjects were trained to the point of mastery.

It might also be the case that there was not enough variation in this artificial language to enable the subjects to form categories. In real language, a listener rarely hears the same speech token more than once. Instead, for any given speech sound, we hear many tokens produced by many different talkers. It is likely that this phoneme-irrelevant information aids in the acquisition of linguistic categories, by highlighting the invariant properties that define the category (see Pisoni 1990, Nygaard et al. 1992).

A third possibility is that learners do not develop featural representations until they have encountered more than one example of the contrast in question. In the case of our experiment, that would require that subjects be trained on the same contrast at *two* places of articulation (e.g. both alveolar and velar) before transfer would occur to a third, untrained place of articulation (e.g. labial).

4. Conclusion

So how far can the input take us? This experiment demonstrated that with input as limited as this artificial language, and only 9 minutes of exposure, subjects do not develop featural representations. The representations that these subjects formed of the trained contrast only enabled them to discriminate the trained contrast itself, and not a second contrast exhibiting the same contrastive feature.

In natural language acquisition, however, learners do appear to represent phonetic and/or phonological categories in terms of contrastive features, and these abstract representations begin to form as early as 9 months of age (Jusczyk et al. 1999). It is possible that the variation inherent in natural language is what enables learners to develop more abstract representations than did the subjects in this experiment. Or perhaps learners don't assign featural structure to phonetic categories until they have good evidence that there are multiple contrasts in the language that exhibit the same feature. This study did not provide a definitive answer to the question of how learners develop abstract, featural representations of speech sounds; however, it lays the groundwork for further investigations into the development of phonological representations.

Endnotes

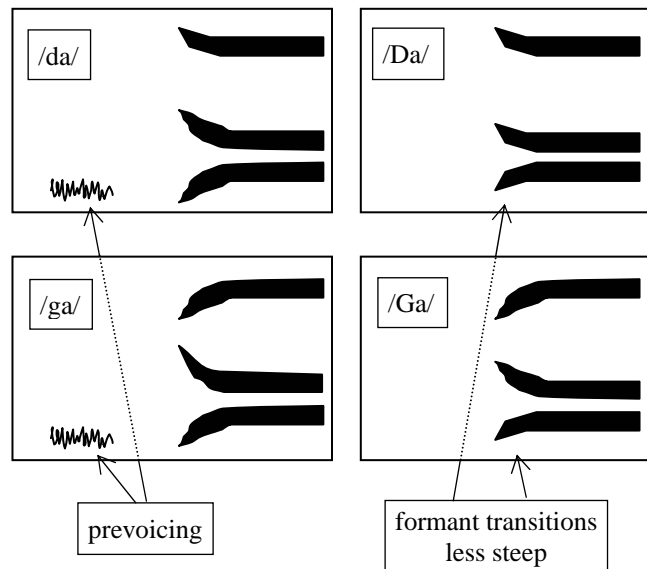
* This research was supported by a University of Arizona SBSRI graduate student research grant to J. Maye, and NSF grant #SBR9696072 to L. Gerken.

1. Syllables beginning with /D/ were created from English syllables beginning with /st/, such as /sta/. After excising the /s/ from these syllables, the remaining unaspirated /t/ sounds like a 'd' to English speakers. To highlight this fact, we have transcribed it as /D/, rather than the customary /t/ of phonetic notation.
2. Despite the resynthesis manipulations on these stimuli, the end-product sounded to naïve listeners like naturally produced speech.
3. To account for speech perception data, a distribution-based model of phonetic category learning must calculate distributions only within the set of sounds occurring in a given phonetic or phonological context. For a complete discussion of this issue, see Maye (2000).

Appendix

Figure 6. Schematic Spectrograms of /da~/~Da/ and /ga~/~Ga/.

References



- Best, Catherine T., Gerald W. McRoberts, & Nomathemba M. Sithole (1988) Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception & Performance* 14, 345-360.
- Jusczyk, Peter W., Mara B. Goodman, and Angela Baumann (1999) Nine-month-olds' attention to sound similarities in syllables. *Journal of Memory and Language* 40, 62-82.

- Kuhl, Patricia K., K. A. Williams, Francisco Lacerda, Kenneth N. Stevens, and Bjorn Lindblom (1992) Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255, 606-608.
- Liberman, Alvin M., Katherine S. Harris, Howard S. Hoffman, and Belver C. Griffith (1957) The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54, 358-368.
- Liu, Zili, and Daphna Weinshall (2000) Mechanisms of generalization in perceptual learning. *Vision Research* 40, 97-109.
- Maye, Jessica (2000) Learning speech sound categories on the basis of distributional information. Unpublished Ph.D. thesis, University of Arizona.
- Maye, Jessica, and LouAnn Gerken (2000) Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development* (p. 522-533). Somerville, MA: Cascadilla Press.
- McClaskey, Christine, David Pisoni, and Thomas D. Carrell (1983) Transfer of training of a new linguistic contrast in voicing. *Perception and Psychophysics* 34, 323-330.
- Nygaard, Lynne C., Mitchell S. Sommers, & David B. Pisoni (1992) Effects of speaking rate and talker variability on the recall of spoken words. *Journal of the Acoustical Society of America*, 91 (4), 2340.
- Pisoni, David B. (1990) Effects of talker variability on speech perception: Implications for current research and theory. *Proceedings of the 1990 International Conference on Spoken Language Processing, Kobe, Japan*, p. 1399-1407.
- Tremblay, Kelly, Nina Kraus, Thomas D. Carrell, and Therese McGee (1997) Central auditory system plasticity: Generalization to novel stimuli following listening training. *Journal of the Acoustical Society of America* 102, 3762-3773.
- Wang, Yue, Michelle M. Spence, Allard Jongman, and Joan A. Sereno (1999) Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America* 106, 3649-3658.
- Werker, Janet F., and Richard C. Tees (1984) Developmental changes across childhood in the perception of nonnative speech sounds. *Canadian Journal of Psychology*, 37, 278-286.