MAPPING FROM ACOUSTIC SIGNAL TO PHONETIC CATEGORY: 
NATURE AND ROLE OF INTERNAL CATEGORY STRUCTURE

Joanne L. Miller
Department of Psychology, Northeastern University, Boston, MA 02115

ABSTRACT

There is now considerable evidence that phonetic categories have a rich internal structure, with some members perceived as better exemplars than others. This paper reports highlights from two investigations that examine this structure and its role in processing. The first investigation demonstrates that contextual factors that have qualitatively similar effects in the region of the category boundary can have qualitatively different effects in the region of the best exemplars of a category. The second investigation demonstrates that the speed of on-line phonetic categorization depends more on an exemplar's position in perceptual space vis-à-vis a boundary with a competing category than on its perceived category goodness. Taken together, these findings underscore the importance of considering both category boundaries and category best exemplars when developing models of speech processing.

1. INTRODUCTION

The early stages of speech perception are often characterized in terms of a perceptual mapping between acoustic signal and prelexical phonetic representations. Although early research focused on the abstract categorical nature of these phonetic categories, more recent findings have shown that phonetic categories are internally structured in a graded fashion, with some members of the category perceived as better exemplars than others (11), (2), (3), (4). In this paper I present highlights from two recent investigations in our laboratory that examine this structure and its role in processing. The first, conducted by Allen and Miller (5), focuses on how different types of contextual factors influence internal category structure. The second, conducted by Miller and Cox (6), focuses on the role of internal category structure in on-line, speeded phonetic categorization.

2. CONTEXT EFFECTS

A ubiquitous finding in the study of speech perception is that the mapping from acoustic signal to phonetic category is highly context-dependent. One form this context dependency takes involves the boundaries between phonetic categories. Numerous studies have shown that both acoustic-phonetic factors and higher-level linguistic factors systematically alter the location of such category boundaries along acoustic continua (7). Over the past few years, we have been investigating whether acoustic-phonetic contextual factors alter not only the location of phonetic category boundaries, but also which stimuli within a category are perceived as the best category exemplars. Our basic strategy for these studies is to create extended speech series that deliberately include poor as well as good exemplars of a category, and ask listeners to rate the stimuli in terms of category goodness. The critical issue is how variation in a given contextual factor affects which stimuli along the acoustic continuum are perceived as the best category exemplars.

For example, in one of our early studies on the acoustic-phonetic contextual factor of speaking rate (8), we created two voice-onset-time (VOT) series. Each series ranged from short VOT values appropriate for /bi/ through longer VOT values appropriate for /pi/ to very long VOT values beyond those appropriate for /pi/. The /p/ with very long VOT values sounded breathy and exaggerated. The two series differed from each other in speaking rate, specified by syllable duration. In a preliminary two-choice identification experiment, we found that as speaking rate slowed, the /bl/-/pl/ category boundary shifted toward longer VOT values, as expected from previous research (9). In the main experiment, which used a category-goodness rating task, we found that the perceived best exemplars of /pi/ also shifted to longer VOT values as speaking rate slowed. In subsequent studies, we have found similar effects on best-exemplar location for other acoustic-phonetic contextual factors, including sentential speaking rate (10) and phonetic-feature context (11) (and see (12) for related findings). Taken together, these studies demonstrate that changes in acoustic-phonetic context can alter the location of the best exemplars of a category along an acoustic continuum.

In the current investigation (5) we asked whether a shift in best-exemplar location would occur if a higher-level linguistic contextual factor, rather than an acoustic-phonetic contextual factor, was manipulated. We focused on the higher-level contextual factor of lexical status, i.e., whether a sequence of phonetic segments forms a word of the language or a nonword. It is well known that lexical status affects category boundary location (13). For example, if listeners are presented stimuli from the two VOT series beef-peef and beace-peace, they identify stimuli with mid-range VOT values along each series so as to form a real word of the language -- beef in the first series and peace in the second series. As a consequence, the /bl/-/pl/ category boundary is located at a longer VOT value for the beef-peef series than the beace-peace series. The question we addressed was whether the effect of lexical status is limited to the region of the category boundary, where by
definition there is ambiguity in category membership, or
whether the effect pervades the category, altering which
stimuli are perceived as the best exemplars.

To answer this question, we used LPC-based speech
synthesis to construct two matched extended VOT series.
Each series ranged from short VOT values appropriate
for /b/ through longer VOT values appropriate for /p/ to
very long VOT values, beyond those appropriate for /p/
The /p/s with very long VOT values (which, for purposes
of explication, we label */p/), sounded breathy and
exaggerated. The VOT series differed from each other
in lexical status. One series ranged from a word to a
nonword (beef-peef-*peef) and the other ranged from a
nonword to a word (beace-peace-*peace). In a
preliminary two-choice identification experiment on
the initial consonant (/b/ vs. /p/), we obtained the expected
effect of lexical status on the /b/-/p/ boundary: the
boundary was located at a reliably longer VOT value for
the beef-peef-*peef series than the beace-peace-*peace
series.

For the main experiment, a different group of
listeners was asked to judge the goodness of the initial
consonant of each token from the two series as an
instance of /p/ using a 1-7 rating scale. We expected to
obtain non-monotonic goodness functions for both
series, with goodness first increasing and then decreasing
as VOT ranged from a short value (more appropriate for
a voiced than a voiceless consonant) to a very long value
(where the voiceless consonant sounded breathy and
exaggerated). The critical question was whether the
stimuli judged to be the best exemplars of the /p/
category would be shifted to longer VOT values for the
beef-peef-*peef series compared to the beace-peace-
*peace series.

Inspection of the group functions (based on smoothed
individual functions) revealed an overall difference in
height between the functions, with the tokens receiving
somewhat higher ratings when the /p/ was judged in the
context of the word (peace) than the nonword (peef). To
adjust for this height difference, which is not relevant to
our main question, we normalized the two group
functions so that they both peaked at a value of 7, and
these normalized functions are shown in the left panel
in Figure 1. The main finding was that the functions are
displaced at their left edge, reflecting the expected
contextual effect in the /b/-/p/ boundary region, but they
converge in the region of the best category exemplars.
Statistical analysis (on the non-normalized data)
confirmed this observation, revealing a reliable shift
between the functions in the boundary region and at the
left edge of the best-exemplar range, but not a reliable
shift at the right edge of the best-exemplar range. (As in
our earlier work, the best-exemplar range was defined,
for each listener, as the range of stimuli receiving ratings
that were at least 90% of the rating given to the highest-
rated stimulus in the series.) That is, a change in lexical
status did not shift the entire best-exemplar range, as
occurs with acoustic-phonetic contextual factors, but
shifted only the left-edge of the range, which borders the
category boundary region.

Figure 1. Normalized group goodness functions
showing the effect of lexical status (left panel) and
speaking rate (right panel) on the internal structure of the
/p/ category.

To ensure that this qualitatively different pattern was
due to lexical status per se, and not to some idiosyncrasy
of our stimuli, we conducted a control experiment in
which we altered the speaking rate of the stimuli in the
beace-peace-*peace series, and tested a new group of
listeners using the same procedures on the fast and slow
versions of the stimuli. The normalized data, which are
displayed in the right panel of Figure 1, show that the
variation in speaking rate did alter the location of the
entire best-exemplar range, as expected from our
previous research, and statistical analysis confirmed this
observation.

These findings suggest a dissociation between the
effect of lexical status and the effect of acoustic-phonetic
factors such as speaking rate on the location of a
category's best exemplars. Acoustic-phonetic factors,
but not the higher-order linguistic factor of lexical status,
alter the location of the entire best-exemplar range of the
category. We suggest that this dissociation derives from the
differential effects of these contextual factors in
speech production. Acoustic-phonetic contextual factors
alter the production of the critical segmental information,
and we propose that listeners track this variation by
altering the location of the perceived best category
exemplars. For example, as speakers slow their rate of
speech, the VOT values of voiceless stop consonants
become longer [11] and, as described above, listeners
correspondingly require longer VOT values to perceive
the best exemplars of slow compared to fast syllables.
In contrast, lexical status does not alter VOT values in
production: in a companion production study [5] we
found no systematic effect of lexical status on the VOT
values of voiceless consonants in words (such as peace)
versus matched nonwords (such as peef). Thus there is
no systematic variation for listeners to track and, as our
perceptual data show, lexical status does not shift the
entire best-exemplar range of the voiceless category in
perception. Note that on this account, although many
contextual factors alter the location of the boundaries
between categories, only those contextual factors that
systematically change the way in which the critical
segmental information is produced will substantially
affec the location of the category’s perceived best exemplars.

3. SPEEDED CATEGORIZATION

The dissociation between different types of contextual factors reported above underscores the importance of considering the internal structure of categories, as well as the boundaries between categories, when studying phonetic perception. Accordingly, in a related set of studies [6], we have been examining how internal category structure, as reflected in category goodness ratings, affects how quickly listeners can map the acoustic signal onto phonetic categories during on-line processing.

For this investigation, we used LPC-based speech synthesis to create nine different extended VOT series, each based on a single word pair. The tokens in three series began with labial consonants (e.g., bat-pat-*pat), the tokens in three series began with alveolar consonants (e.g., den-ten-*ten), and the tokens in three series began with velar consonants (e.g., ghost-coast-*coast). In a preliminary category-goodness rating task, in which listeners were asked to rate each token for the goodness of its initial voiceless stop consonant (/p/, /t/, or /k/) using a 1-7 rating scale, we obtained the expected non-monotonic goodness function for each series. Specifically, category goodness first increased and then decreased as VOT ranged from a short value (more appropriate for a voiced than a voiceless consonant) to a very long value (where the voiceless consonant sounded breathy and exaggerated).

For the first of two main experiments, we selected tokens from each series that covered a wide range of VOT values and varied considerably in category goodness. We presented these tokens to three groups of listeners. Each group heard the tokens from the three series with a given place of articulation randomized together. The listeners’ task was to decide as quickly as possible whether each word began with /b/ or /p/ (labial group), /d/ or /t/ (alveolar group) or /g/ or /k/ (velar group) and to indicate their response by pressing one of two labeled keys (e.g., B or P). The main question was how quickly listeners could perform this two-choice categorization. On the basis of the literature [14], we expected that response time would be relatively slow for the poor voiceless category exemplars near the voiceless boundary, and relatively fast for the best voiceless category exemplars. At issue was the response time for poor voiceless category exemplars with VOT values longer than those of the best exemplars. If response time is a direct function of category goodness, then response time for these stimuli should be relatively slow. However, if response time depends more on the position of the stimulus vis-à-vis the category boundary, then response time should be relatively fast in that these stimuli are far from the voiceless boundary and are clearly identified as members of the voiceless category.

Overall, similar findings were obtained for all nine stimulus series. Representative data are shown for one of the series, den-ten-*ten, in the left panel of Figure 2, where the response time function for the relevant stimuli (those ranging from the voiced-voiceless boundary region to the longest VOT values) is presented along with the goodness function for the series. The main finding across the nine series, confirmed by statistical analysis, was that response time was relatively slow for poor voiceless exemplars near the voiceless-voiceless category boundary, but relatively fast for poor voiceless exemplars with long VOT values – overall, poor voiceless exemplars with long VOT values were categorized as quickly as the best exemplars themselves.

These findings provide initial support for the view that categorization time is more dependent on position vis-à-vis the category boundary than on perceived category goodness. However, note that in the experiment, listeners in a given group made two-choice speeded categorization judgments on tokens that all had a single place of articulation. Given this design, it is possible that listeners were not fully analyzing the poor exemplars with long VOT values, but simply quickly categorizing them as voiceless by default, due to their breathy, exaggerated quality. To control for this possibility, we conducted a second main experiment with new listeners using the same stimuli, but a task that involved monitoring as well as categorization. Again, three groups of listeners were tested. However, this time each group heard the stimuli from all nine series randomized together. The listeners’ task was to monitor for two initial consonants, /b/ or /p/ (labial group), /d/ or /t/ (alveolar group), or /g/ or /k/ (velar group). When they heard one of the target consonants they were to respond as quickly as possible by pressing one of two labeled keys (e.g., B or P); if the word did not begin with one of the target consonants, they were to refrain from responding. With this task, listeners could not use a default strategy of simply pressing the key corresponding to the voiceless consonant whenever they heard a poor exemplar that was breathy and exaggerated, as the task required listeners to analyze the stimuli for place of articulation as well as voicing. The question was whether with this new task, we would also find that

![Figure 2](image-url)
poor exemplars of the voiceless category with long VOT values were categorized quickly.

The answer was yes. Overall, response times were longer than in the first experiment (as to be expected, given the additional monitoring task), but the same general pattern prevailed. The data for the den-ten—ten series are shown in the right panel of Figure 2. Statistical analysis across the nine series confirmed that, as in the first experiment, response time was relatively slow for poor voiceless exemplars near the voice-d-voiceless category boundary, but relatively fast for poor voiceless exemplars with long VOT values—overall, poor exemplars with long VOT values were categorized as quickly as the best exemplars themselves. Taken together, the data from the two tasks demonstrate that categorization time is not a direct function of perceived category goodness. Instead, the time it takes to map the critical acoustic information onto a phonetic category appears to depend more on the exemplar's location in perceptual space vis-à-vis a boundary with a competing category.

4. CONCLUSIONS

The findings described in this paper provide additional support for the view that phonetic categories have a rich internal structure, and they provide new information on the nature of this structure and its role in processing. First, the findings indicate that although some contextual factors that alter the location of a boundary between phonetic categories along an acoustic continuum also alter the location of a category's best exemplars, other contextual factors have a more limited effect, primarily confined to the boundary region. Second, they indicate that the speed with which listeners map the acoustic signal onto a given phonetic category depends more on the exemplar's position in perceptual space vis-à-vis a boundary with a competing category than on its perceived category goodness. Taken together, these findings underscore the importance of considering both category boundaries and category best exemplars when developing models of the early stages of speech processing.

5. ACKNOWLEDGEMENTS

This research was supported by NIH Grant DC 00130 to J.L. Miller and by a Minority Fellowship from the Acoustical Society of America to Sean Allen. The author thanks Sean Allen and Larry Brancacio for their helpful comments on an earlier version of the paper. Correspondence should be addressed to: Joanne L. Miller, Department of Psychology, Northeastern University, Boston, MA 02115. E-mail: jim@neu.edu.

6. REFERENCES